

خوانغ تشانغ نينغ
لي جوان تزي
علم الذخائر اللغوية
ترجمة: هشام موسى المالكى
المركز القومى للترجمة
2523



إن العمل الذي بين يدي القارئ هو ترجمة من اللغة الصينية لمُؤَلَّف عن علم الذخائر اللغوية. يُعْتَبَر علم الذخائر اللغوية من العلوم اللغوية التأسيسية التي تُرَسِّخْ مفهوم دراسة اللغة في بيئتها الطبيعية، بعيداً عن القياس اللغوي المنطقي الذي ساد في حقل الدراسات اللغوية قروناً عدة. إن علم الذخائر اللغوية علم يبحث في كيفية جمع النصوص اللغوية الطبيعية وتهيئتها وترميزها؛ بحيث تكون صالحة للبحث اللغوي ودراسة الظواهر اللغوية الطبيعية على مستوى أفرع علم اللغة بنظرياته وتطبيقاته الحديثة. يحتل علم الذخائر اللغوية، باعتباره أحد المنهجيات التي تمهد لدراسة اللغة الطبيعية بشكل موضوعي، مكانة متقدمة في حقل اللسانيات الحديثة. وقد استُثْمِرَت منهجية الذخائر اللغوية في الأبحاث اللغوية التي تخدم علم اللغة التطبيقي بأفرعه المختلفة وفي مقدمتها اللغويات الحـــــاسوبية، وصناعة المعاجم، وعلم المصطلح، وتعليم اللغات لأهلها أو للأجانب، والترجمة تعليماً وممارسة، وما إلى ذلك. والجدير بالذكر أن هذا العلم قد نما وتطور تحت مظلة علم اللغة الحاسوبي.

تصميم الغلاف: رحاب حسن

علم الذخائر اللغوية

المركز القومى للترجمة
تأسس فى أكتوبر ٢٠٠٦ تحت إشراف: جابر عصفور
مدير المركز: أنور مغيث

- العدد: 2523
- علم الذخائر اللغوية
- خوانغ تشانغ نينغ، ولى جوان تزى
- هشام موسى المالكى
- اللغة: الصينية
- الطبعة الأولى 2016

هذه ترجمة كتاب:
语料库语言学
黄昌宁　李涓子 著

شارع الجبلاية بالأوبرا- الجزيرة- القاهرة.　ت: ٢٧٣٥٤٥٢٤　فاكس: ٢٧٣٥٤٥٥٤
El Gabalaya St. Opera House, El Gezira, Cairo.
E-mail: nctegypt@nctegypt.org　Tel: 27354524　Fax: 27354554

علـــم
الذخائر اللغوية

تأليـــف: خوانـــغ تشانغ نينـــغ
لى جـــوان تـــزى
تـــرجمـــة: هشـــام موسى المـــالكى



2016

بطاقة الفهرسة
إعداد الهيئة العامة لدار الكتب والوثائق القومية
إدارة الشئون الفنية

نينغ، خوانغ تشانغ

علم الذخائر اللغوية/ تأليف: خوانغ تـشانغ نينـــغ، لـــى جوان تزى، ترجمة: هشام موسى المالكى.

ط١، القاهرة، المركز القومى للترجمة، ٢٠١٦

٥٤٠ ص، ٢٤سم

١- اللغة العربية - تاريخ.

(أ) تزى، لى جوان (مؤلف مشارك)

(ب) المالكى، هشام موسى (مترجم)

(ج) العنـــوان ٤١٠,٩

رقم الإيداع : ٢٥٠٩٩ / ٢٠١٤

الترقيم الدولى : 978-977-92-0000-2

طبع بالهيئة العامة لشئون المطابع الأميرية

المحتويات

كلمة المترجم

يُعْتَبَر علم الذخائر اللغوية[1] من العلوم اللغوية التأسيسية التي تُرَسِّخ مفهوم دراسة اللغة في بيئتها الطبيعية، بعيدًا عن القياس اللغوي المنطقي الذي ساد في حقل الدراسات اللغوية قرونًا عدة. إن علم الذخائر اللغوية، الذي أسَّسَ له عالم اللغة الإنجليزي ليتش (Leech)[2] في النصف الثاني من القرن العشرين، هو علم يبحث في كيفية جمع النصوص اللغوية الطبيعية وتهيئتها وترميزها؛ بحيث تكون صالحة للبحث اللغوي ودراسة الظواهر اللغوية الطبيعية على مستوى أفرع علم اللغة بنظرياته وتطبيقاته الحديثة. يحتل علم الذخائر اللغوية- باعتباره أحد المنهجيات التي تمهد لدراسة اللغة الطبيعية بشكل موضوعي- مكانة متقدمة في حقل اللسانيات الحديثة. ولا غنى للباحث اللغوي عن التعرف على مفاهيم هذا العلم وفنياته وتطبيقاته، بل وطُرُق بناء الذخائر اللغوية بأحجامها المختلفة لخدمة أغراض بحثية معينة.

---

(١) لمزيد من المعلومات عن تأصيل مصطلح "الذخائر اللغوية"، انظر:
هشام موسى المالكي، "إشكاليات تهيئة الذخائر اللغوية وبنائها حاسوبيًا -اللغتان العربية والصينية- نموذجًا"، مجلة أواصر، المجلد الثاني، المركز القومي للترجمة، القاهرة، ٢٠٠٩، ص ٢٨-٥٦.

(٢) جيفري ليتش (Geoffry Leech)، من مواليد ١٩٣٦، عمل أستاذًا للغويات واللغة الإنجليزية الحديثة بجامعة لانكاستر البريطانية في الفترة من ١٩٧٤ حتى ٢٠٠٢، ويعمل أستاذًا متفرغا بقسم اللغويات واللغة الإنجليزية بالجامعة نفسها منذ ٢٠٠٢ حتى الآن، كما أنه عضو بالأكاديمية النرويجية للعلوم والآداب.

فالذخيرة اللغوية، وفقًا لمفاهيم علم الذخائر اللغوية، هي بناء لغوي يتمتع بمواصفات ومعايير فنية تجعله قادرًا على استيعاب النصوص اللغوية وإتاحتها للبحث اللغوي العام والخاص. وينبغي أن تخضع النصوص اللغوية التي يحتويها هذا البناء إلى قواعد معينة من حيث أساليب الجمع، ونِسَب التمثيل، وطرق المعالجة قبل عملية الجمع وبعدها، ومنهجيات الترميز والأساليب التي يتم على أساسها عمليات الاستعلام والاستدعاء حسب متطلبات البحث اللغوي.

وقد استُثْمِرَت منهجية الذخائر اللغوية في الأبحاث اللغوية التي تخدم علم اللغة التطبيقي بأفرعه المختلفة وفي مقدمتها اللغويات الحاسوبية، وصناعة المعاجم، وعلم المصطلح، وتعليم اللغات لأهلها أو للأجانب، والترجمة تعليمًا وممارسة، وما إلى ذلك. والجدير بالذكر أن هذا العلم قد نما وتطور تحت مظلة علم اللغة الحاسوبي. وقد تفاعل العِلْمَين معًا أخذًا وعطاءً على مدى ما يزيد عن نصف قرن؛ حيث ساهم علم الذخائر اللغوية بما يُقَدِّمُه من مفاهيم ومنهجيات في تطوير آليات علم اللغة الحاسوبي في التعامل مع المادة اللغوية الطبيعية موضع البحث سواء على المستوى الشفهي أو المستوى التحريري. وعلى الجانب الآخر، فقد ساهم تطور علم اللغة الحاسوبي الذي تواكب مع تطور إمكانات الحاسوب في تقديم الأدوات التقنية التي كان لها أكبر الأثر في تقدم مفاهيم ونُظُم بناء الذخائر اللغوية. وعلى هذا الأساس، يمكننا القول بوجود علاقة طردية بين المجالين، تجعل كلا منهما يؤثر في الآخر دفعًا وتطويرًا.

وعلى الرغم من ذلك، فإنني أنظر إلى هذا العلم بمفاهيمه، وآلياته، ونظرياته من خلال منظور أوسع يجعل فائدة علم الذخائر اللغوية تخرج من إطار اللغويات الحاسوبية لتشمل البحث العلمي في كل ما يتصل بالإنتاج اللغوي عند أي جماعة لغوية سواء في اتجاه أحادي، أو في اتجاه تقابلي أو مقارن مع لغات أخرى. وأخُصُّ بذلك مجالات الإبداع الأدبي وبخاصة ما يتعلق بالنقد الأدبي وتحليل النصوص الأدبية؛ حيث تُمَكِّن أدوات هذا العلم الباحث من جمع النصوص الإبداعية بأنواعها المختلفة وتهيئتها للبحث العلمي، وتطوير آليات البحث فيها عن طريق ترميزها بمنهجيات مختلفة تتيح الفرصة لإجراء عمليات استعلام متعددة الوظائف عليها لاستخراج الظواهر الأدبية، والخصائص الأسلوبية للنصوص موضع التحليل أو النقد. كما يمكن أن تمتد فائدته لدراسة النصوص الدينية وتفسيراتها أو ترجماتها المختلفة بغرض الوقوف على أوجه التشابه والاختلاف بين التفاسير أو الترجمة، أو جمع النصوص التراثية وتحليلها من وجهة نظر علم اللغة التداولي، وتحقيقها وما إلى ذلك من قضايا بحثية. وتحقيق تلك القضايا البحثية بالطرق التقليدية لا يتيح الوصول إلى النتائج المرجوة إلا بعد استنزاف وقت وجهد طائلين من الباحثين، بالإضافة إلى تحجيم نطاق النتائج التي يتم تحقيقها، فضلاً عن عدم وجود تراكم للنصوص المدروسة، التي لا تتم معالجتها وفقًا لمعايير علم الذخائر اللغوية بما يُمَكِّن من الاستعانة بها في مستويات بحثية أخرى.

والعمل الذي بين يدي القارئ هو ترجمة من اللغة الصينية لمؤلَّف عن علم الذخائر اللغوية الذي ترجع أصوله إلى اللغة الإنجليزية، ثم أخذ رحلته

في التطور والارتقاء عبر لغات مختلفة سواء من حيث التأليف وعقد سلاسل المؤتمرات الدورية المتخصصة، أو من حيث مشروعات الذخائر اللغوية الضخمة التي تُنشئها الدول وترعاها وتمولها بأموال طائلة؛ تمهيدًا لطرحها للاستخدام على مستويات بحثية مختلفة. وهنا قد يتبادر إلى الذهن سؤال عن الدافع وراء ترجمة هذا العلم عن اللغة الصينية؟ فما دام أن علم الذخائر اللغوية قد نُقِلَ إلى الصين عن اللغة الإنجليزية، فلماذا لا ننقل عن اللغة الإنجليزية مباشرة؟ والإجابة على هذا السؤال ألخصها في النقطتين التاليتين:

أولاً: على الرغم من كثرة المتعاملين باللغة الإنجليزية واللغات الهندأوروبية والمترجمين عنها، ففي حدود المعلومات المتوفرة لدي، لم يتقدَّم أحد بطرح فكرة ترجمة كتاب عن هذا العلم عن أية لغة من اللغات. ويرجع السبب في ذلك إلى عدم وجود قسم علمي متخصص للغويات الحاسوبية في الجامعات المصرية، مع ندرة الكوادر المتخصصة في هذا المجال في مصر، والذين تتلقفهم الدول الأجنبية فيذهبون للبحث في الجامعات الأجنبية هَرَبًا من البيئة المحلية غير المُهَيَّئة للعمل العلمي المنضبط في مجال اللغويات الحاسوبية؛ الأمر الذي أدى إلى وجود حالة من التضارب المفهومي والاصطلاحي في مؤلفاتنا اللغوية الحديثة فيما يتعلق بمنهجيات هذا العلم وأهميته وطرق توظيفه في البحث اللغوي التطبيقي. ومن وجهة نظري، يُعْتَبَر هذا الوضع أحد أسباب تراجع حال البحث اللغوي في اللغة العربية بالمفاهيم الحديثة لعلم اللغة، وتَوَجُّه اللغويين لدينا إلى اتخاذ موقف الحاكي عن تجارب الدول فيما يتعلق بعلم اللغة الحديث، بدلاً من تَصَدُّر المشهد،

وتولي مسئولية الفعل والتطبيق؛ وذلك باستثناء عدد قليل من الباحثين الذين يعملون بشكل منفرد. وذلك على خلاف ما يحدث على صعيد اللغات الأخرى. بالإضافة إلى ذلك، تنعدم التطبيقات اللغوية النابعة من احتياجاتنا الواقعية، الأمر الذي أدى إلى حالة من التبعية الدائمة للشركات الأجنبية في كل ما يتعلق بالتطبيقات اللغوية التقنية التي تتعامل مع اللغة العربية سواء على مستوى مشروعات البرمجيات الحاسوبية التي تتعامل مع اللغة الطبيعية، أو على مستوى التعامل المتقدم باللغة العربية على محركات البحث عبر الشبكة، وما إلى ذلك من تطبيقات حققت نتائج غير مسبوقة في لغات أخرى.

ثانيًا: إن النقل عن اللغة الصينية ليس كما يتبادر إلى ذهن العامة من أن الصين هي تلك الدولة البعيدة التي تنتمي إلى دول العالم الثالث، التي تُصدِّر لنا "فوانيس رمضان" والسلع رخيصة الثمن "المضروبة"، أو تلك الدولة التي تمتلك لغة صعبة مفرداتها عبارة عن مجموعة هائلة من الرسوم توقفت عن النمو وملاحقة ركب التطور اللغوي منذ مئات القرون، وما إلى ذلك من مفاهيم مغلوطة مرجعها إلى توجهنا الدائم بأنظارنا إلى الغرب، وتمحور اهتمامنا في التعلم من ذلك المكان الذي تغرب فيه الشمس، علمًا بأن الشمس تشرق دائمًا من الشرق. إن اللغة الصينية هي أثمن ما تملكه تلك الدولة البعيدة الغنية بمواردها البشرية التي تجعل تجربة النقل عنهم تجربة ثرية على الرغم من صعوبتها بل استحالتها في بعض الأحيان إذا ما قورنت بلغات أخرى كالإنجليزية. ولكن ما يزيد التجربة ثراءً هو نقل ذلك الفهم لعلم الذخائر اللغوية لدى الصينيين الذين نجحوا بالفعل في معالجة قضايا شائكة

أصعب ما تكون مقارنة بما تعانيه اللغات الأخرى. فقد كانت تجربتهم في معالجة اللغة الصينية اعتمادًا على الذخائر اللغوية في المقام الأول، ثم انطلقت المراكز البحثية والجامعات بقوافل من الباحثين وأعداد هائلة من المؤتمرات المتخصصة، وفي النهاية توصلوا إلى توطين مفهوم علم الذخائر اللغوية. الأمر الذي قادهم إلى التوصل إلى حلول جذرية لمشكلات المعالجة الآلية للغة الصينية الطبيعية على جميع الأصعدة على الرغم من صعوبة لغتهم؛ ولذلك أرى أن تجربة النقل عن اللغة الصينية فيها من الثراء بما لا يمكن أن يوجد إذا نقلت عن الإنجليزية صاحبة الفكرة الأساسية.

فالصينيون نجحوا في توطين علم الذخائر اللغوية وتوظيفه بشكل عملي، وأنتجوا من خلاله التطبيقات الحاسوبية التي تُعَبِّر عن هويتهم الذاتية في حل الإشكاليات ذات الخصوصية اللغوية، وعلى رأسها التمييز الآلي لحدود الكلمات في النصوص التحريرية المكتوبة باللغة الصينية، أو التمييز الآلي للغة التحريرية سواء المكتوبة بحروف مطبعية أو بخط اليد وأنظمة إدخال اللغة إلى الحاسب الآلي، أو إدارة الاستعلامات المتقدمة باللغة الصينية عبر الشبكة، أو التواصل بين العنصر البشري والآلة، وما إلى ذلك؛ الأمر الذي مكَّن الصينيين من إضافة بصمة واضحة المعالم في الوعاء المعرفي لعلم الذخائر اللغوية.

ومن هنا تولَّدَ لديَّ الدافع الأول لمبادرة توطين هذا العلم في اللغة العربية. مما جعلني أتحمس لمبادرة ترجمة هذا الكتاب عن اللغة الصينية بعد ما شاهدته في أثناء تعاملي مع الصينيين العاملين في مجال اللغويات

الحاسوبية في رحلات علمية مختلفة إلى الصين من اهتمام بحثي ومتابعة لأحدث التطورات العالمية في علم الذخائر اللغوية. بالإضافة إلى أن اللغة الصينية هي اللغة التي أتخصص في الترجمة منها وليست الإنجليزية، فضلاً عن السبب الذي سبق ذكره من عدم مبادرة العاملين في الترجمة عن الإنجليزية أو اللغات الأخرى وما أكثرهم بتقديم هذا العلم وترجمة أحد هذه المؤلفات وتوطين مفهوم علم الذخائر اللغوية في الوعاء المعرفي للغة العربية. وهذا الكتاب هو الكتاب الأول الذي تتم ترجمته إلى اللغة العربية في مجال علم الذخائر اللغوية، على أمل أن تؤدي تجربة الترجمة إلى نقل المزيد من المؤلفات عبر لغات أخرى، وإلى التأليف باللغة العربية. وقد رأيت بنفسي أنه بقدر اهتمام الدول بعلم الذخائر اللغوية وتطوير آلياته، بقدر تطور البحث اللغوي التطبيقي للغاتها. وبقدر الفهم المتعمق لأي عالم من علماء اللغويات لمفاهيم هذا العلم ومنهجياته، وآلياته، بقدر تعاظُم ما يمكن أن يصل إليه من نتائج تطبيقية. وبقدر ما تهتم به الدول، بقدر ما ترقى لغاتها إلى مستوى الدخول في عصر المعلومات.

والمُؤلِّف الرئيس لهذا الكتاب هو الأستاذ الدكتور خوانغ تشانغ نينغ، أحد أشهر علماء اللغة الحاسوبيين في الصين. وُلِدَ خوانغ تشانغ نينغ عام ١٩٣٧ في مقاطعة جوانغ دونغ، وفي عام ١٩٦١ تخرج في قسم المحركات الكهربية بجامعة تشينغ خوا ببكين. وفي عام ١٩٨٣ سافر في بعثة إلى جامعة ييل (Yale·University) بالولايات المتحدة الأميركية لمدة عام، وعمل في التدريس لمدة عام بجامعة العلوم والتكنولوجيا بهونج كونج عام ١٩٨٦. وفي عام ١٩٩٩، عمل مديرًا لمركز أبحاث شركة ميكروسوفت في الصين.

ويُنظَر إليه الصينيون باعتباره ممهد الطريق لتوطين علوم المعالجة الآلية للغات الطبيعية في الصين. في سبعينيات القرن العشرين، ذاع صيته كأستاذ متميز في جامعة تشينغ خوا عندما أَلَّفَ مُقَرَّر ((دوائر الترانزستور))؛ حيث حقق هذا المقرر مبيعات زادت عن مليون نسخة داخل الصين. وبعد ذلك، قَدَّمَ ترجمة لكتابين أحدثا أثرًا كبيرًا في المجتمع العلمي الصيني هما كتاب ((مبادئ الذكاء الاصطناعي))، وكتاب ((البرمجة باستخدام لغة **LISP**))، ليملأ بذلك الفراغ العلمي الموجود في تلك الفترة.

وقد وصل خوانغ تشانغ نينغ إلى قمة أخرى من قمم الشهرة العلمية عندما عمل في حقل الدراسات المتعلقة بعلم معالجة اللغات الطبيعية؛ حيث بدأ بتنظيم مجموعات بحثية داخل الصين وخارجها في هذا المجال، ثم تولى رئاسة العديد من المؤتمرات الدولية وهيئات المبرمجين في مجال المعالجة الآلية للغات الطبيعية. ويتولي خوانغ تشانغ نينغ الآن رئاسة تحرير ((المجلة العلمية للمعلوماتية الصينية))، بالإضافة إلى كونه مُحكمًا وعضو هيئة تحرير العديد من المجلات العلمية الدولية مثل ((المجلة التخصصية في معالجة المعلومات باللغات الآسيوية **ACM**)) بالولايات المتحدة الأميركية، و((النشرة الدورية لجمعية معالجة المعلومات باللغة الصينية واللغات الشرقية)) بسنغافورة، ومجلة ((معالجة اللغات الطبيعية)) باليابان، ومجلة ((دراسات نقدية حول علم اللغة في الصين)) بهونج كونج، ومجلة ((علم اللغة الحاسوبي ومعالجة اللغات الصينية)) بتايوان[١].

---

(١) لمزيد من المعلومات عن المؤلف، انظر الموسوعة الصينية على الرابط التالي:
http://baike.baidu.com/view/1615502.htm

وقد شارك خوانغ تشانغ نينغ في مشروعين بحثيين في إطار الخطة الخمسية السابعة للحكومة الصينية عام ١٩٨٥، هي: مشروع "الفهم الآلي للغات الطبيعية والبرامج البينية بين الإنسان والآلة" وكان ذلك في نطاق مشروعات العصف الذهني، ومشروع "تقنيات الفهم الآلي للنصوص العسكرية" في نطاق المشروعات البحثية المستقبلية للدفاع الوطني، بالإضافة إلى مشروع "نظام للترجمة الآلية بين اللغتين الصينية واليابانية" بالتعاون بين جامعتي تشينغ خوا ونانكين الصينيتين. وقد حصل هذا المشروع على المركز الثاني في مجال تطوير العلوم والتكنولوجيا من لجنة التعليم بالحكومة الصينية.

وقد سبق أن تولى خوانغ تشانغ نينغ رئاسة الفريق البحثي الصيني الذي كُلِّفَ من وزارة الصناعات الإلكترونية الصينية للمشاركة في مشروع الترجمة الآلية متعدد اللغات، الذي رصدت له الحكومة اليابانية خمسين مليون دولار أميركي عام ١٩٩٠ لتطوير نظام للترجمة الآلية بين لغات خمس دول آسيوية هي: اليابان، والصين، وماليزيا، وسنغافورة، وتايلاند.

وفي عام ١٩٩٥ شارك في مشروع بحثي للترجمة الآلية بين اللغتين الإنجليزية والصينية ممثلا عن جامعة تشينغ خوا الصينية بالتعاون مع شركة DEC الأميركية. كما نجح مع فريقه البحثي في بيع حقوق الملكية الفكرية لعدد من النتائج البحثية لشركتي آي بي إم IBM، وميكروسوفت **Microsoft**.

ينقسم كتاب ((علم الذخائر اللغوية)) إلى خمسة أبواب وملحق للاختصارات وقائمة مراجع.

الباب الأول: عبارة عن مقدمة تعريفية بالذخائر اللغوية، وماهية علم الذخائر اللغوية، وتاريخ تطور العلم وآفاقه المستقبلية، والدور الذي يلعبه الحاسب الآلي في علم الذخائر اللغوية، والموضوعات البحثية التي يهتم بها هذا العلم.

الباب الثاني: يتحدث عن تصميم الذخائر اللغوية وتطويرها، من حيث بعض مشكلات جمع المادة اللغوية، وكيفية بناء ذخيرة لغوية، وأهم الموضوعات التي ينبغي أخذها بعين الاعتبار حتى تكون الذخيرة مُمَثِّلة للواقع اللغوي، بالإضافة إلى عرض لأنواع الذخائر اللغوية، وتعريف بأشهر الذخائر اللغوية على مستوى العالم والصين.

الباب الثالث: يتطرق إلى معالجة الذخائر اللغوية والتقنيات المستخدمة في إدارتها، وطرق إدارة الاستعلام داخل الذخائر اللغوية، وتوظيف علم الإحصاء في علم الذخائر اللغوية، وأشهر برمجيات الفهرسة المعروفة مثل برمجية كولوكيت Collocate، وتيبيكال Typical، بالإضافة إلى ترميز الذخائر اللغوية من حيث المفهوم والغرض، وأنواع الترميز المختلفة للنصوص اللغوية على كل من المستويات النحوية والدلالية، والترميز على مستوى الخطاب.

الباب الرابع: يتحدث عن العلاقة بين علم الذخائر اللغوية وعلم اللغة، من حيث استخدام منهجية الذخائر اللغوية في الدراسات اللغوية على مستوى البحوث المتعلقة بعلم المفردات، والنحو، والدلالة، بالإضافة إلى علم اللغة التداولي وتحليل الخطاب، فضلاً عن تعليم اللغات وعلم المعجمية.

الباب الخامس: يتناول تطبيقات منهجية الذخائر اللغوية في علم اللغة الحاسوبي، من خلال دراسة الأبحاث المتعلقة بإزالة اللبس الدلالي، والدراسات المتعلقة بتمييز التعبيرات اللغوية الاسمية وما إلى ذلك.

لم تكن تجربة نقل كتاب ((علم الذخائر اللغوية)) عن اللغة الصينية تجربة سهلة، لعدة أسباب أهمها عدم وجود وعاء معرفي سابق باللغة العربية في الترجمة والتأليف في هذا المجال، الأمر الذي جعل رحلة الترجمة بكل ما تحتويه من مصطلحات وما تشير إليه تلك المصطلحات من مفاهيم علمية أمرًا شاقًا ومحفوفًا بالمخاطر؛ خشية الإخفاق في التعبير عن المصطلح بلغة مناسبة سهلة الفهم، فضلاً عن أن أي خلل في الصياغة من شأنه أن يؤدي إلى خطأ في التوجيه لا يستطيع المترجم تحمل عواقبه العلمية. لذلك أنتهز هذه الفرصة للتأكيد على ضرورة المراجعة ومداومة الفعل وتوظيف ما ورد بالكتاب من مصطلحات ومفاهيم عن طريق الممارسة واستمرار الترجمة، لدفع المصطلحات والمفاهيم الواردة بالكتاب إلى مزيد من السلاسة والتداول. ومن ضمن أسباب الصعوبة أيضًا أن النقل قد تم عن اللغة الصينية وما يُحتِّمُه هذا النقل من مواجهة بعض المشكلات الخاصة باللغة الصينية، التي ليس لها وجود في لغات أخرى وبخاصة اللغة العربية، ما أدى إلى استغراق الكتاب وقتًا مضاعفًا عما كان مخططًا له في البداية، إلى أن تمكنت من تدشين المصطلحات الجوهرية لهذا العلم الوليد في اللغة العربية. وقد ساعدني في التغلب على هذه المشكلات عدد من القراءات الإثرائية باللغتين الصينية والإنجليزية، فضلاً عن ممارسة الإنتاج العلمي والبحثي في بعض القضايا

التي يتناولها الكتاب بالتوازي مع تقدم العمل في الترجمـة. فقـد كـان إذن لتخصصي في الترجمة واللغويات الحاسوبية عظيم الأثر في سبر أغوار علم الذخائر اللغوية، وتسهيل معالجة بعـض المفـاهيم المُركّبـة والمنهجيـات المُستَغلَقة، وصياغة المصطلحات المعَبّرَة عنها بأكبر قدر من الوضوح؛ حتى يتم توسيدها في الوعاء المعرفي للغة العربية؛ إيمانًا من المتـرجم بأهميـة العلوم البينية وتداخل التخصصات اللغوية (كاللغويـات الحاسـوبية وعلـم الذخائر اللغوية) في العمل على تكامل البحث اللغوي التطبيقي.

وسعيًا إلى توسيع دائرة فائدة القارئ العربي من الكتاب قدمت بعـض الإضافات في النسخة المترجمة تتمثل فيما يلي:

١- مقدمة بقلم المترجم للتعريف بأهمية الكتاب والمؤلف.

٢- عدد من الحواشي استعنت بها لشرح بعـض المفـاهيم الخاصـة بطبيعة اللغة الصينية أو التعريف ببعض العلماء الذين لهم إسهامات في مجال الذخائر اللغوية أو اللغويات الحاسوبية.

٣- فهرس باللغتين الصينية والعربية يحتـوي قائمـة المـصطلحات الواردة في الكتاب الأصلي مع إضافة المعنى المقابل بالإنجليزيـة لعدد من المصطلحات الأساسية في هذا التخصص، لإفادة الباحثين والمترجمين العاملين في هذا المجال، ولتكون بداية لتـوطين تلـك المصطلحات في الوعاء المعرفي للغة العربية، من خلال مداومـة العمل على تطويرها وتنميتها واستخدامها بين المتخصصين.

٤- ترجمة للمراجع الصينية التي استعان بها مؤلف الكتاب لتعريف القارئ بمستوى ما توصل إليه الصينيون في هذا المجال.

٥- ترجمة للأمثلة التطبيقية الصينية الموجودة في الكتاب وقد تم وضعها دائمًا بين علامتي تنصيص، لزيادة مفهومية النص المترجم.

وفئة القراء المستهدفة من ترجمة كتاب "علم الذخائر اللغوية" هم العلماء والباحثون المتخصصون في مجال اللغويات الحاسوبية على وجه الخصوص، والعلماء والباحثون في حقل اللغويات التطبيقية بصورة عامة، وخاصة في مجالات علم المعجمية، وعلم المصطلح، والترجمة، وتعليم اللغات، وما إلى ذلك. كما يستهدف الكتاب المتخصصين في الدراسات الأدبية؛ حيث يمكنهم الاستعانة بمنهجيات هذا العلم في دراساتهم النقدية والأسلوبية اعتمادًا على تحليل ذخائر لغوية أدبية. ويمكن اعتبار هذا الكتاب مرجعًا للباحثين المشتغلين في بحوث لغوية تطبيقية، أو بحوث أدبية تستعير من اللغويات أدوات التحليل والعرض؛ حيث يساعد هذا الكتاب على ترسيم مُخَطَّط علمي منضبط يمكن أن يَعْتَمِد عليه الباحثون في عملهم.

وأود التأكيد على أن هذا العمل يُعَدُّ بالنسبة إليَّ نقطة البداية لسلسلة من الأعمال المترجمة التي أحلم بتوطينها في اللغة العربية عن علم اللغة الحاسوبي بقضاياه وتوجهاته البحثية المختلفة. إن ترجمة كتاب "علم الذخائر اللغوية" بالنسبة إلى هذه الأعمال هو الخطوة الأولى نحو توطين تلك العلوم

باللغة العربية. على أمل أن يحظى هذا الجهد بقبول حسن، وأن يُسْهِمَ في تعريف القارئ العربي بعلم الذخائر اللغوية وكيفية الاستفادة منه من قبل الباحثين في بناء ذخائر لغوية عامة وخاصة تساعد على دراسة اللغة العربية في الاتجاه التطبيقي والتحليق بها إلى آفاق التطبيق، حتى تخرج هذه اللغة العريقة من سُباتها البحثي والعملي الذي تسبب فيه أهلها، وتنحصر الفجوة بيننا وبين من سبقونا، وحتى تعود الشمس لتشرق من الشرق.

أ.د./ هشام موسى المالكي

أستاذ الترجمة واللغويات الحاسوبية

قسم اللغة الصينية

كلية الألسن- جامعة عين شمس

القاهرة - أبريل - ٢٠١٣م

الباب الأول

مقدمة ومدخل إلى الموضوع

أصبح "علم الذخائر اللغوية (Corpus Linguistics) بالفعل فرعًا مــن فروع علم اللغة. وإجراء الأبحاث القائمة على الذخائر اللغوية لم يعد مقتصرًا على خبراء علوم الحاسب فقط، بل أصبح لهذا العلم تأثيرٌ متزايدٌ على جميع المجالات البحثية في علم اللغة". هذه العبارة ذُكِرَت في مقدمة الكتاب القــيِّم الذي ألَّفَه توماس وآخرون عام ١٩٩٦ بمناسبة الاحتفال بمرور ستين عامًــا على ميلاد عالِم اللغة ليتش (Leech)(١) مؤسس علم الذخائر اللغويـــة. وقــد تزايدت الإشارات والنصوص المعبرة عن المضمون نفـــسه فـــي الـــسنوات الأخيرة في المؤلفات اللغوية الشهيرة التي تتناول منهجية العمل في البحوث اللغوية وأساليبها، وأصبح هذا المصطلح من المصطلحات المتعارف عليهـــا لدى جميع العاملين في علوم اللغة[1].

(١) جيفري ليتش (Geoffrey Leech) من مواليد ١٩٣٦، عمل أستاذًا للغويات واللغة الإنجليزية الحديثة بجامعة لانكاستر البريطانية في الفترة مـــن ١٩٧٤ حتـــى ٢٠٠٢، ويعمل أستاذًا متفرغًا بقسم اللغويات واللغة الإنجليزية بالجامعة نفسها منذ ٢٠٠٢ حتى الآن، كما أنه عضو بالأكاديمية النرويجية للعلوم والآداب. (المترجم)

الفصل الأول

الذخائر اللغوية وعلم الذخائر اللغوية

يُقصَد بالذخيرة (Corpus) ذلك الكيان (أو قاعدة البيانات) الذي تُخَزَّن بداخله النصوص اللغوية. وقد جرى العرف بين علماء اللغة أن مصطلح ذخيرة لغوية يُقصَد به ذلك العدد الهائل من الشواهد اللغوية الطبيعية الذي يُستَخدم كأساسٍ لإجراء البحوث اللغوية. والذخيرة اللغوية لا تقتصر فقط على اللغة التحريرية؛ ولكنها تحتوي أيضًا على عينات صوتية من اللغة الشفهية، وقد تقتصر الذخيرة على أحد مستويات اللغة فقط، سواء المستوى الكتابي أو المستوى الشفهي، هذا بالإضافة إلى أن هذا المفهوم يُستَخدم أيضًا للدلالة على اللغة في حالة ثابتة، أو اللغة في حالاتها المتغيرة على حدٍّ سواء. وقبل اختراع الحاسب الآلي كان الباحثون -على وجه الخصوص مؤلفو المعاجم اللغوية – يمتلكون ذخائر لغوية، إلا أن نطاق تلك الذخائر كان محدودًا للغاية؛ ولذلك كان من الصعب تكوين تيار علمي يمثل هذا المصطلح في الأوساط اللغوية. ولكن منذ حوالي أربعين عامًا، أصبح هذا المصطلح يقترن دائمًا بالمادة اللغوية الطبيعية التي تُخَزَّن عن طريق الحاسب الآلي، التي تُستَخدم في الدراسات اللغوية والتطبيقات الخاصة بهندسة اللغة. ومع تطور الحاسب وتقنياته تطورت الذخائر اللغوية وتعددت أشكالها وتضاعف حجمها؛ بالإضافة إلى تنوع أغراض استخدامها وتطبيقاتها. وفي الوقت ذاته

ظهرت أعداد كبيرة من البرمجيات المتخصصة التي تُستَخْدم في حفظ نصوص الذخائر اللغوية واسترجاع المعلومات من تلك النصوص ومعالجتها. وتطورت الذخائر اللغوية الإلكترونية بصورة سريعة لتصبح مصدرًا مهمًّا من مصادر البحث اللغوي. وقد تأسست العديد من الذخائر اللغوية فائقة الحجم على مستوى العالم، منها ما هو عالمي، ومنها ما هو محلي على مستوى الجامعات ومؤلفي ودور نشر المعاجم. هذا بالإضافة إلى أن التطور السريع للحاسبات الشخصية وانخفاض تكلفة وسائط تخزين البيانات قد أتاح الفرصة أمام كل باحث لتكوين ذخائر لغوية تتلاءم مع احتياجاته وميوله البحثية.

وعلى الرغم من أن هناك تاريخًا ملموسًا لعلم الذخائر اللغوية، فإنه لم يتم الاتفاق على تعريف محدد لهذا العلم حتى الآن. وفيما يلي نستخلص تعريفين لعلم الذخائر اللغوية:

التعريف الأول: علم الذخائر اللغوية **(Corpus linguistics)** هو ذلك العلم الذي يهتم بدراسة اللغة على أساس من أمثلة لغوية واقعية تُستَمد من واقع الحياة الطبيعية للجماعة اللغوية. (ماكنري وولسون ١٩٩٦) [2].

التعريف الثاني: علم الذخائر اللغوية هو ذلك العلم الذي ينطلق من الوصف اللغوي للنصوص اللغوية أو يتخذ من الشواهد اللغوية الواقعية داخل النصوص نقطة انطلاق لتحديد فرضيات البحث اللغوي. (كريستال ١٩٩١) [3].

ويمكننا أن نلاحظ من خلال التعريفين السابقين أن مصطلح علم "الذخائر اللغوية" باعتباره مصطلحًا علميًّا، يختلف اختلافًا جوهريًّا عن "علم النحو"، و"علم الدلالة" في كونه لا يهتم بالنظر إلى جانب واحد من جوانب

اللغة؛ ولكنه منهجية بحث لغوية تعتمد على نصوص اللغـة الطبيعيـة فـي دراسة الظواهر اللغوية. وفي الحقيقة أن هذا العلم يضم توجهين أساسيين:

أولاً: تهيئة اللغة الطبيعية وترميزها لخدمة البحوث اللغوية.

ثانيًا: إجراء الدراسات اللغوية وتطوير التطبيقات اللغوية على الذخائر بعد تهيئتها وترميزها.

الفصل الثاني

تاريخ تطور علم الذخائر اللغوية

يرجع تاريخ علم الذخائر اللغوية –باعتباره أحد المنهجيات التي تَدْرُس اللغة الطبيعية– إلى القرن الماضي، وتَرْجِع الوثائق[1] التي ذكرت ذلك العلم إلى فترة تَكوُّن نظرية النحو التحويلي التوليدي لناعوم تشومسكي، ومنذ ذلك الوقت حتى الآن يمكننا تقسيم تاريخ ذلك العلم إلى ثلاث مراحل[1]:

١ــ البدايات الأولى لعلم الذخائر اللغوية (corpus linguistics)

كان ذلك قبل النصف الأول من خمسينيات القرن العشرين، ويُقْصَد بذلك جميع البحوث اللغوية التي كانت تنطلق من المواد اللغوية الطبيعية في فترة ما قبل ظهور نظرية النحو التحويلي التوليدي لتشومسكي، ويمكن تمثيل التطبيقات التي نتجت في تلك المرحلة من خلال النقاط التالية:

أ) اكتساب اللغة

هذا الحقل من أول المجالات التي استُخْدِمت فيه الذخائر اللغوية كوسيلة من وسائل البحث اللغوي. ففي سبعينيات القرن التاسع عشر، ظهر في أوروبا أول تيار من الدراسات يبحث في طرق اكتساب الأطفال للغة. وفي الوقت ذاته، كانت جميع المواد اللغوية موضع الدراسة هي تلك

اليوميات التي تُدوِّن حديثَ الوالدين للأبناء وتَطَوُّر هذا الحديث مع تقدم الأبناء في السن. وكما هو معروف كانت تلك اليوميات هي المادة الأولية التي اعتُبِرَت أحد المصادر اللغوية التي بنى عليها العلماء فرضياتهم اللغوية أمثال بريار Preayer [4] وستيرن Stern [5] وغيرهما، بالإضافة إلى أنها ما زالت تُعتَبر أحد المصادر اللغوية التي يعتمد عليها الباحثون حتى يومنا هذا. ومنذ ثلاثينيات القرن العشرين، قدم علماء اللغة وعلماء النفس العديد من الأنماط اللغوية للأطفال في أعمار مختلفة. وقد وُضِعَت تلك الأنماط اللغوية بصورة أساسية اعتمادًا على كميات هائلة من المواد اللغوية التي تُسَجِّل اللغة الطبيعية للأطفال.

ب) علم اللهجات

ارتبط علم اللهجات منذ نشأته ارتباطًا وثيقًا بالذخائر اللغوية، وقد وُلِد هذا العلم من رحم علم اللغة التاريخي المقارن الذي نشأ في القرن التاسع عشر، وكانت الأسباب الأولى لنشأة هذا العلم تنحصر في استخدام الوسائل المباشرة لتسجيل الاختلافات الصوتية داخل اللغة لعرض خريطة اللهجات في لغةٍ ما. "وكان الباحثون في البداية يحملون في أيديهم كراسات لتدوين ملاحظاتهم، ثم تلى ذلك استخدام أجهزة التسجيل الصوتي، وكانوا يُدَوِّنون أو يُسَجِّلون كل ما يعترضهم من مواد لغوية تُعبِّر عن اختلاف اللهجات. وهذه المادة ما زالت مرجعًا علميًّا لبعض الباحثين غير المتفرغين، ولها قيمة واضحة في مجال دراسة توزيع المفردات داخل اللهجات". (francis,1980)[6].

ج) تعليم اللغات

يُعتَبر كلٌ من بـونجر Bonger، وترافـر Traver، وفرايـز fries (١٩٤٧) هم أول من استخدم الذخائر اللغوية في دراسة أساليب تعليم اللغـة وتعلمها. وهذا يتطابق مع ما ذكره كينيدي Kennedy (١٩٩٢)[7] من وجود ارتباط وثيق بين الذخائر اللغوية وتعليم اللغات الأجنبية في الخمـسين سـنة الأولى من القرن العشرين. وقال: إن المصدر الـرئيس لقـوائم المفـردات المستخدمة في تعليم اللغات الأجنبية كان تلك الإحصاءات التي تمـت علـى الذخائر اللغوية في اللغات موضع الدراسة. وكان لتلك القوائم فائدة كبرى في توجيه عملية تعليم اللغات الأجنبية.

د) النحو والدلالة

قام بعض علماء اللغة بتوصيف اللغة اعتمادًا على الذخائر اللغوية مثل عالم اللغة فرايز (Fries) (١٩٥٢) الذي قدم وصفًا لقواعد اللغة الإنجليزيـة قياسًا على دراسة ذخيرة لغوية [8]. وهذه المحاولة تسبق ما قام به عالم اللغة كويرك (Quirk) وغيره في أواخر الثمانينيات لتأليف "موسوعة قواعد اللغة الإنجليزية" بحوالي ثلاثين عامًا.

هـ) دراسة النظام الصوتي للغة

يعود استخدام ذخائر اللغات الطبيعية في دراسة النظام الصوتي للغـة إلى علماء اللغة البنيويين أمثال فريدريك بـواس F.Boas، وإدوارد سـابير E.Sapir وغيرهما، فقد اهتموا بالعمل في البيئة الطبيعية للغة وأكدوا علـى

ضرورة أن ينطلق البحث اللغوي من اللغة الطبيعية ومن النظرة الموضوعية لها. وهذه الآراء كانت حجر الأساس الذي أرسى دعائم علم الذخائر اللغوية. ومع حلول الفترة الأولى من خمسينيات القرن العشرين أصبح المناخ العلمي مُهيئًا للاهتمام بدور الذخائر اللغوية في دراسة اللغة. الأمر الذي شكل تيارًا قويًا في الوقت ذاته وخاصة في أميركا، على يد زيليك هاريس (Harris) وغيره من علماء اللغة البنيويين خلفاء ليونارد بلومفيلد؛ حيث اعتبروا أن الذخائر اللغوية هي المادة الأساسية التي يجب أن تنطلق منها دراسات اللغة، وكانت رؤيتهم تنصب على أن الشواهد اللغوية التي تقوم على الحدث يجب أن تأتي في المرتبة الثانية وأنها لا يُعتمد عليها في بناء النظام اللغوي ويجب أن تُستَبْدل بقواعد تنطلق أساسًا من الواقع اللغوي.

٢ـ مرحلة النحو التحويلي التوليدي لتشومسكي

بعد ظهور النظرية النحوية لتشومسكي [9] عام ١٩٥٧ وما تلاها من سلسلة أبحاث لغوية، حدث تغير جذري في تاريخ تطور علم الذخائر اللغوية الذي تحدثنا عنه سالفًا. وتحول التيار الأساسي لمنهجيات البحث اللغوي تبعًا لذلك من النزعة التجريبية (Empiricism)(١) إلى النزعة العقلية (Rationalism)(٢)، في دراسة اللغة. ففي تلك الفترة، احتلت نظرية ديكارت

---

(١) النزعة التجريبية: مذهب من يقيم المعرفة على ما تدركه الحواس وحدها، وينكر وجود مبادئ فطرية في النفس وقوانين صادرة عن العقل. (المترجم)

(٢) النزعة العقلية: مذهب في الفلسفة يرى أن كل شيء في الوجود مرده إلى العقل، ويتخذ من العقل والاستنباط معيارًا للحقيقة بدلا من المعايير الحسية. وقد أخذ بهذا المذهب ديكارت وهيجل وليبنز في فلسفتهم. (المترجم)

العقلية مكان الصدارة، وتراجعت النزعة التجريبية لتصبح كما لو أن ليس لها أساس تستند إليه. وبطبيعة الأحوال أُنكرت جميع أنواع الذخائر اللغوية التي تكونت على أساس من النزعة التجريبية. ويعتقد أصحاب المذهب العقلي أن جزءًا كبيرًا من المعلومات اللغوية لدى الإنسان توقيفية موروثة. أما أصحاب المذهب التجريبي فيقفون على النقيض من هذا الرأي؛ حيث يعتقدون أن اللغة اصطلاحية، وأن معلومات الإنسان تدخل إلى عقله أولاً عن طريق الحواس، ثم تتكون بعد ذلك المعلومات اللغوية عن طريق بعض التصورات البسيطة وبعض المعالجات التي تتم على تلك التصورات الذهنية، وأن الإنسان عندما يولد لا يكون متمتعًا بالمبادئ اللغوية وأساليب معالجة اللغة. ونظرًا إلى استقبال نظرية تشومسكي استقبالا حسنًا في الأوساط اللغوية، فقد هيمنت النظرية العقلية على الدراسات اللغوية على مدى ستينيات وسبعينيات القرن الماضي ولمدة ما يقرب من عشرين عامًا.

وقد تركزت الانتقادات التي وجَّهَها تشومسكي وزملاؤه من علماء النحو التحويلي التوليدي للمحاولات الأولى للبحث اللغوي باستخدام الذخائر اللغوية على النقاط التالية:

أ- وجود أخطاء في أساليب البحث اللغوي المبني على الذخائر اللغوية. فقد فرّق تشومسكي بين القدرة اللغوية (language competence) ، والأداء اللغوي (language performance)؛ حيث يُعتَقَد أن الهدف الأساسي من البحث اللغوي يقوم بشكل رئيس على القدرة على وضع أنماط للمعرفة اللغوية تُعَبِّر عن الواقع

الداخلي لنفس المتحدث، بمعنى تقديم أنماط القدرة اللغوية، لأننــا لا نستطيع أن نقدم وصفًا وشرحًا للمعلومات اللغوية لدى المتحدث إلا عن طريق القدرة اللغوية. أما الاســتخدام اللغــوي فهــو الــدليل الظاهري على القدرة اللغوية، وهو دائمًا ما يتخطى حــدود تــأثير التحولات التي تحدث على العناصر الأساسية المُكَوِّنة للغة. ومــن هنا فلا يمكن أن يكون الاستخدام اللغوي معبرًا بصورة كاملة عن القدرة اللغوية. ويعتقد تشومسكي أيضًا أن جوهر المعلومات اللغوية ما هو إلا مجموع العناصر المُكَوِّنة للغة؛ وأن أنماط التجريب التي تمخضت عن دراسة الذخائر اللغوية لا تقدم إلا تفسيرًا جزئيًّا للقدرة اللغوية. ومن هنا، نجد أن هذا الأسلوب في دراسة اللغة لا يتمتــع بالقوة الكافية.

ب– عدم كفاية المادة اللغوية؛ حيث اكتشف تشومسكي لأول مرة فــي كتاب (النظرية النحوية) أن قواعد تركيب التعبيــرات فــي اللغــة الإنجليزية تتسم بالاضطراد، وهذا الاضطراد يُعَبِّر عن أن الجمــل في اللغات الطبيعية ليس لها حدود؛ وبمعنى آخر أن المواد اللغوية ليس لها نهاية ولن تصل إلى حد الاكتمال أبدًا. وقد غَيَّــرَت هــذه الرؤية لعلماء اللغة التحويليين اتجاه البحث اللغوي الذي ساد طَوَال خمسينيات القرن الماضي عند علماء اللغة البنيويين. وفيما يقــرب من عشرين سنة بعد ذلك، يمكننا القول: إن جميع علماء اللغة قــد اتبعوا ذلك النهج في دراسة اللغة. وساد أسلوب استخدام الــذخائر

اللغوية وفقًا للرؤية الجديدة في بحث اللغة بصورة هائلــة. ولكــن على الرغم من ذلك فإن البحث في علم الذخائر اللغوية لم يــصل إلى نهايته، بل ظهر العديد من العلماء الذين اتسموا بشجاعة علمية غير عادية وتحملوا أعباءً لا حصر لها في ســبيل تطــوير علــم الذخائر اللغوية. ففي عام ١٩٥٩، قام عالم اللغة كويرك (Quirk) ببناء ذخيرة لغوية بعنوان: (دراسة مسحيَّة لطرق اســتخدام اللغــة الإنجليزية). وفي الوقت ذاته، بدأ كل مــن فرانــسيز (Francis)، وكوسيرا (Kucera) العمل في مشروعهم الضخم لبنــاء الــذخيرة اللغوية الشهيرة التي عُرِفَت باسم براون Brown التــي اســتغرق بناؤها ما يقرب من عشرين عامًا حتى اكتملت تمامًا، وبالإضــافة إلى ذلك بدأ عالم اللغة جان سفارتفيك (Jan Svartvik) فــي عــام ١٩٧٥ بناء ذخيرة لندن النصية المعروفة باســم ذخيــرة لنــدن-لوند (London-Lund Corpus) علــى أســاس مــن الــذخيرتين اللغويتين السابقتين، بالإضافة إلى أنه توصل أخيرًا إلى بناء ذخيرة لغوية مُمَيْكَنة عُرِفَت باسم SEU. وبناءً على تلك الجهود رأى ليتش (Leech:1991) أن: "مصادر دراسة اللغة الــشفهية الإنجليزيــة لا يمكن منافستها حتى الآن"، وبالإضافة إلى ذلك نجد أن التعاون الذي تم بين خبراء الحاسب وعالمي اللغة فرانسيز وكوسيرا في جامعــة براون قد تمخض عنه بناء أول ذخيرة لغوية يقرأها الحاسب فــي عام ١٩٦١. وقد استخدمت تلك الذخيرة اللغوية عينات من اللغــة

الإنجليزية تحت موضوعات مختلفة، ووصل نطاق تغطيتها إلى حوالي مليون كلمة بغرض دراسة النموذج الأميركي من اللغة الإنجليزية. وتُعْتَبر هذه الذخيرة اللغوية بمثابة حجر الأساس لعلم الذخائر اللغوية الذي تأسس في ثمانينيات القرن الماضي.

٣ـ فترة ازدهار علم الذخائر اللغوية وانطلاقه

منذ ثمانينيات القرن الماضي وعلى مدى ما يقرب من عشرين عامًا، تطور علم الذخائر اللغوية بصورة كبيرة ونما نموًّا متسارعًا، وقد تمثل ذلك في النواحي التالية:

أ) توالي ظهور الجيل الثاني من الذخائر اللغوية

فمنذ ثمانينيات القرن الماضي، ظهرت مجموعة كبيرة من الذخائر اللغوية أشهرها الذخيرة المعروفة باسم كوبويلد **Cobuild**، وعلى الرغم من كون تلك الذخائر اللغوية مختلفة فيما بينها من حيث الحجم وأسلوب التصميم والهدف، فإن غالبيتها كان يعتمد على تقنية التمييز الكهروضوئي للحروف والمعروفة بتقنية **KDEM**؛[1] ومن ثَمَّ، تمكن العلماء من تحويل النصوص آليًا إلى الشكل الرقمي وتحريرها من الأسلوب اليدوي الشاق الذي يعتمد كتابة النصوص باستخدام لوحة المفاتيح. إن الذخائر اللغوية التي ظهرت في تلك الفترة لم تتضاعف من حيث الحجم فحسب، بل زادت سرعة بنائها؛ ومن ثَمَّ اصطُلِح على تسمية تلك الذخائر بالجيل الثاني من الذخائر اللغوية. وتشير

---

(1) Korowai Data Entry Machine.

دراسة إحصائية غير كاملة للعالم اللغوي الأميركي إدواردز (**Edwards**) من جامعة كاليفورنيا – بركلي نشرَها عام ١٩٩٣ أن عدد الذخائر اللغوية التي طُرِحَت للاستخدام في ثمانينيات القرن الماضي قد زاد عن خمسين ذخيرة، ويمكن تقسيمها حسب اللغة كما يلي:

الإنجليزية ← 24	**الفرنسية ← 4**	**الإيطالية ← 2**	**الدنماركية ← 2**
الألمانية ← 4	**الإسبانية ← 2**	**الفنلندية ← 2**	**السويدية ← 7**

وبالإضافة إلى ذلك، نجد أن لغات مثل البرتغالية واليوغوسلافية وغير ذلك من اللغات قد تتابعت في إنشاء ذخائر لغوية. ومن بين أعداد الذخائر سالفة الذكر نعرض بالتفصيل بعضًا منها اتسم بالتميز وكِبَر الحجم، كما يلي:

(١) ذخيرة لانكاستر– أوسلو – برجن

(Lancaster – Oslo – Bergen Corpus)

تسمى هذه الذخيرة اختصارًا باسم ذخيرة لوب **LOB**، وقد أسسها فريق عمل تحت قيادة العالم اللغوي ليتش **leech** بجامعة لانكستر الإنجليزية الشهيرة بغرض دراسة اللغة الإنجليزية. وقد تم تصميم هذه الذخيرة باستخدام أسلوب مشابه لأسلوب بناء ذخيرة براون من حيث تقسيم المادة، وأسلوب انتقاء العينات. وقد تم الانتهاء من العمل في هذه الذخيرة عام ١٩٨٣. وتتكون هذه الذخيرة من حوالي خمسمائة عينة لغوية، يصل حجم كل عينة إلى ألفي كلمة. وبذلك يصل إجمالي حجم النصوص التي تحتويها تلك الذخيرة إلى مليون كلمة، ونظرًا إلى تلك المميزات فدائمًا ما يعتبر العلماء ذخيرتي لوب وبراون النصيَّتَيْن بمثابة أختين من أصل واحد يمكن الاعتماد عليهما في مقارنة النموذج البريطاني والنموذج الأميركي من اللغة الإنجليزية.

(٢) ذخيرة النصوص الفرنسية المعروفة باسم TLF

(Trésor de la Langue Française)

أُنشِئَت هذه الذخيرة بالتعاون بين كل من مركز البحوث العلمية التابع للحكومة الفرنسية وجامعة شيكاغو الأميركية، وقد غطت المادة اللغوية في تلك الذخيرة اللغوية الفترة من القرن السابع الميلادي حتى القرن العشرين؛ حيث ضمت ما يقرب من ألفي عينة نصية تغطي جميع الأساليب التحريرية في اللغة الفرنسية. ويصل عدد الكلمات في تلك الذخيرة إلى حوالي مائة مليون وخمسمائة ألف كلمة. وقد حُفِظَت تلك الذخيرة النصية في أسطوانات ليزر بالإضافة إلى استخدام نظام التشغيل يونكس Unix للتعامل معها وتصفحها.

(٣) ذخيرة هيلسنكي النصية للمعلومات التاريخية

(The Helsinki Corpus of Historical English)

تأسست تلك الذخيرة في جامعة هيلسنكي من خلال مجموعة من علماء اللغة بقيادة العالم اللغوي روسيني (Roseanne)، وتشتمل تلك الذخيرة النصية على جميع النصوص التاريخية باللغة الإنجليزية في الفترة من عام ٨٥٠م حتى عام ١٧٢٠م. وقد قُسمت المادة اللغوية بهذه الذخيرة إلى شرائح زمنية مدى كل منها مائة عام. ويصل حجم تلك الذخيرة إلى ست عشرة مليون كلمة. وتُعْتَبر أول ذخيرة نصية تاريخية باللغة الإنجليزية، ولهذه الذخيرة اللغوية قيمة علمية كبيرة من حيث كونها موجَّهة لدراسة مدى التغير في اللغة الإنجليزية من خلال وجهة نظر علم اللغة الاجتماعي وعلم اللهجات والاستخدام اللغوي.

(٤) الذخيرة الدولية للنصوص الإنجليزية ICE

(The International Corpus Of English)

تولى إنشاء هذه الذخيرة قسم اللغة الإنجليزية بجامعة لندن في عام ١٩٨٨، وتهدف إلى تقديم معلومات للأبحاث المقارنة عن معدل التغير اللغوي في الدول الناطقة بالإنجليزية على مستوى العالم، وقد اتخذت عيناتها اللغوية من جميع الدول الناطقة بالإنجليزية، كما استخدمت برمجيات موحدة للتصنيف والتكويد. وكانت حدود المادة اللغوية من كل دولة حوالي مليون كلمة مقسمة بالتساوي بين المستويين التحريري والشفهي. واستغرق جمع محتويات تلك الذخيرة النصية ثلاث سنوات من ١٩٩٠ إلى ١٩٩٣. بالإضافة إلى أن نطاق المادة اللغوية كان يستهدف الأشخاص البالغين الذين يتجاوز عمرهم ١٨ عامًا وتربوا على اللغة الإنجليزية منذ صغرهم.

والجدير بالذكر أن مجموعات العمل المُكلَّفة بإنشاء ذخائر لغوية بالصين كانت تتزايد بصورة متوالية، كما أن الأبحاث اللغوية والمسح اللغوي القائم على ذخائر نصية كان يتزايد باضطراد، مثل الدراسات الإحصائية التي تمت على ذخائر لغوية كبيرة الحجم للتعرف على معدل تكرار الرموز والكلمات في اللغة الصينية. وقد خرج عن تلك الدراسات نتائج علمية ملموسة، من أهمها (قائمة الرموز الأكثر استخداما في اللغة الصينية الحديثة)، و(معجم معدل تكرار الكلمات في اللغة الصينية الحديثة). وسوف نتحدث بالتفصيل عن بناء الأنواع المختلفة من الذخائر اللغوية في الصين بصورة أكثر تفصيلاً في الباب الأخير من هذا الكتاب.

ب- زيادة الأبحاث العلمية القائمة على ذخائر لغوية

لقد أدى إنشاء الذخائر اللغوية إلى دفع الأبحاث اللغوية القائمة على تلك الذخائر، ويمكننا بوضوح أن نلحظ ذلك من خلال الإحصاءات المبيَّنَة في الجدول التالي:

جدول ١-١: إحصاء بالأبحاث القائمة على ذخائر لغوية في الفترة من ١٩٥٩ إلى ١٩٩١

جوهانسون (Johansson:1991)

الفترة الزمنية	عدد الأبحاث
1965 - 1959	10
1970 - 1966	20
1975 - 1971	30
1980 - 1976	80
1985 - 1981	160
1991 - 1986	320

وتشير الحقائق إلى أن الذخائر اللغوية الإلكترونية قد قدمت مصادر رائعة وسَّعَت مجالات البحث اللغوي. وذلك نظرًا إلى أن المادة اللغوية التي تقدمها تتصف بالواقعية بالإضافة إلى وضوح التكوين الهيكلي لها؛ ومن ثَمَّ أصبحت تلك الذخائر خير مُعين على تقديم وصف لغوي متعدد المستويات، بالإضافة إلى أنها ساهمت في تطوير الأبحاث المقارنة بين الأساليب اللغوية

المختلفة، والأكثر من ذلك أنها أسهمت أيضًا في عمل الدراسات الإحصائية الكَمِّية عن اللغة.

وفي تلك الفترة، توصلت العديد من الأبحاث اللغوية إلى نتائج علمية مهمة، منها ما تمكن من تعميق نتائج الدراسات الموجودة بالفعل، ومنها أيضًا ما وسَّع مجالات البحث اللغوي المتعارف عليها. مثل دراسات نحو الاحتمالات التي قدمها هاليداي (**Halliday**) عام ١٩٩١[10]، وسفارتفيك (**Svartvik**) عام ١٩٩٢[11]، وغيرهما. وكذلك الدراسات المتعلقة بخصائص اللغة الإنجليزية التي قدمها دوتي **Dotti** عام ١٩٩١، بالإضافة إلى الدراسات الإحصائية التي قدمها سينكلي (**Sinclair**) وغيره عام ١٩٨٥ عن تراكيب اللغة الإنجليزية.

أما عن أسباب ازدهار علم الذخائر اللغوية في اللغة الإنجليزية مع بداية ثمانينيات القرن العشرين فهناك العديد من الآراء نستطيع أن نوجزها في النقطتين التاليتين:

(١) إن تطور علوم الحاسب وانتشار التقنيات والتطبيقات الإلكترونية قد أرسى دعائم وأُسُسًا قوية لتطور علم الذخائر اللغوية ونموه. فمنذ ثمانينيات القرن الماضي، دخلت الذخائر اللغوية دورة ممتازة من أطوار التطور؛ فقد ساهم النمو الهائل في سرعة الحاسب وقدرته على تخزين المعلومات في الإسراع ببناء الذخائر اللغوية، بالإضافة إلى زيادة القدرة على معالجتها، وتعدد مستويات المعالجة. ومن جهة أخرى، نجد أن القدر الهائل من البيانات التي

تم ترميزها ومعالجتها قد دعَّم أبحاث الذخائر اللغويــة وطوَّرهــا وعظَّم من فرص الاستفادة منها. وفي تلك الفترة، ظهــرت أيــضًا وسائل بحثية متطورة ولغات برمجة ساهمت في ابتكار برمجيــات حديثة حوَّلت مجرى الدراسات الإحصائية ومعالجــة اللغــة مــن الأسلوب اليدوي إلى أسلوب مُمَيْكَن أو نصف مُمَيْكَن. ومن هنا يبدو لنا جَلِيًّا الدور المهم للحاسب في ذلك التطور والنمو.

(٢) ثَبَتَ أن بعض النقد الذي وجهه علماء مدرسة النحــو التحــويلي التوليدي لعلم الذخائر اللغوية كان خاطئًا، مثل الرأي الــذي يــتهم تقنية الحاسب أنها تقنية خادعة، بالإضافة إلى أن عددًا آخر من تلك الآراء النقدية كان سطحيًّا، على سبيل المثال ذلك الرأي الداعي إلى الإنكار الكلي لقيمة الذخائر اللغوية؛ وبعضه كان صــحيحًا مثــل الرأي القائل: إن عدد الجمل التي تتكون في إطار اللغة لا متناهي. أما بالنسبة إلى الرأي الداعي لاتخاذ المذهب العقلي فــي دراســة اللغة، فقد أثبت تدريجيًّا بعض القصور بعد أن اتبعه العديــد مــن العلماء وطبقوه وأعملوا فيه فكرهم، وتَمَثَّل ذلك القصور في العجز عن التحقق من الفرضيات اللغوية العقلية التــي يــضعها العلمــاء وإثباتها وغير ذلك. ولذلك فقد عبر علم الذخائر اللغوية منذ نــشأته في ثمانينيات القرن العشرين بصورة واسعة عن أن مجال البحــث في علم اللغة متسع وشاسع. ويعني ذلك اســتعادة التــوازن بــين المعلومات الطبيعية والمعلومات التوقيفية في الدراســات اللغويــة.

فلماذا لا يتم الجمع والتكامل بين العنصرين في البحث اللغوي حتى يمكن الاستفادة من مميزات كل منهما. ولكي يمكن الوصول إلى ذلك التوازن النافع دعا العديد من علماء اللغة إلى ما يلي:

"المادة اللغوية تلعب دورًا محوريًا في النظريات اللغوية، وهناك العديد من المسارات التي يمكن اتباعها لتطوير تلك المادة... ولا يمكن القول بوجود طريقة واحدة". هاليداي (**Halliday:1991**).

"من وجهة النظر العلمية، يمكن القول: إن أسلوب الذخائر اللغوية هو الأسلوب الأقوى في دراسة اللغة، نظرًا إلى أن النتائج التي يتم التوصل إليها يمكن قياسها والتحقق منها". ليتش (**Leech:1993**).

حتى إننا نجد عالمًا مثل فيلمور (**Fillmore**)، والذي وجه انتقادات كبيرة إلى علم الذخائر اللغوية، يكتب ما يحمل معنى الاعتراف بفضل الذخائر اللغوية على العلوم اللغوية: "علَّني أعتقد عدم وجود الذخيرة النصية التي تقدم معلومات كاملة عن قواعد اللغة الإنجليزية ومفردات... ولكن في الحقيقة أنه في كل مرة أطَّلع على نتائج معالجة لذخيرة نصية مهما كان حجمها أجد أنها قدمت ما لا يمكن أن تقدمه أية طريقة أخرى. وخلاصة قولي في هذا الأمر هو أن كلا المذهبين يحتاج كل منهما إلى الآخر". فيلمور (**Fillmore:1992**).

الفصل الثالث

توجهات التطوير وآفاقه في علم الذخائر اللغوية

تحدث العديد من العلماء المتخصصين في علم الذخائر اللغوية في السنوات الأخيرة عن آفاق التطور في ذلك العلم بصورة مستفيضة، مثل سفارتفيك الذي تنبأ قائلاً: "سوف يصاحب ذلك ازدياد في حجم الذخائر اللغوية وتحسُّن في جودتها، وسوف يرتفع معدل استخدامها" **(Svartvik:1992)**.

أما ماكنري، فعلى العكس من ذلك يرى أن تطور علم الذخائر اللغوية سوف يتأثر تطوره بأربعة عوامل هي: نطاق الذخائر اللغوية وأنواعها ومدى اهتمام الدول بها، ومعدل تطور الحاسبات الآلية **(McEnery:1996)**. أما بالنسبة إلى الوضع الراهن للأبحاث اللغوية القائمة على الذخائر اللغوية، فقد وجدنا أن الآراء حول توجهات التطور في ذلك العلم سوف تتمثل في النواحي التالية:

١- تطور الاعتماد على الذخائر اللغوية

مع بداية تسعينيات القرن العشرين تزايد سعي الدول للاستثمار في بناء الذخائر اللغوية؛ نظرًا إلى زيادة الوعي بقيمة المصادر اللغوية القومية، وظهر ذلك السعي بصورة أكبر بين دول أوروبا انطلاقًا من رغبة الحكومات

والهيئات العلمية بتلك الدول في الحفاظ على لغاتها القومية، مثل جامعة أكسفورد الإنجليزية التي أصدرت الذخيرة اللغوية القومية بالتعاون مع جامعة لانكستر ومؤسسة رومانس الإنجليزية للطبع والنشر والمكتبة الملكية الإنجليزية وغير ذلك من الهيئات، ومثل هيئة الثقافة والعلوم والتربية اليابانية التي كونت مجموعة عمل من ثلاثمائة خبير عام ١٩٨٩ قضوا خمسة أعوام في بناء الذخيرة اللغوية للغة اليابانية، وقد سُجِّلت هذه الذخيرة على عدد ٢٢ أسطوانة ليزر، من بينها عدد ٩ أسطوانات لغوية وعدد ٣ أسطوانات للمعلومات. وسوف يستمر البحث في بناء مثل هذا النوع من الذخائر اللغوية. ويأتي ترميز الذخائر اللغوية في المرحلة التالية بعد إنشائها. وهناك من يعتقد أن السنوات المقبلة سوف تُسفِر عن ظهور ذخائر لغوية صغيرة الحجم لأغراض محددة، مثل الذخيرة اللغوية التي أنشأها شورت لدراسة أساليب التعبير عن اللغة والفكر (Short:1996). وبالإضافة إلى ذلك فإن الذخائر اللغوية المبنية على اللغة الشفهية ينبغي أن تتطور بصورة أسرع حتى يمكن التغلب على حالة عدم التوافق بين الذخائر اللغوية التحريرية والذخائر الشفهية، ومن أجل دفع الدراسات المتعلقة باللغة الشفهية. وقد توقع ولسون أنه في المستقبل القريب سوف نجد ذخائر نصية متعددة الوسائط (Wilson:1996).

٢ـ تطور أساليب ترميز المواد اللغوية

يُقصد بترميز الذخائر اللغوية وضع أساس لإجراء تحليل متعدد الأبعاد والمستويات، بالإضافة إلى أن المستفيد النهائي من ذلك الترميز لا يكون بالضرورة هو الشخص الذي قام بعملية الترميز في البداية؛ ومن ثَمَّ نستطيع القول: إن الاستخدام الأمثل للذخائر اللغوية يعتمد بدرجة كبيرة على

مستويات الترميز وجودته داخل تلك الذخائر. ولذلك فإن القوة المُحَرِّكة لتطور ترميز الذخائر اللغوية يجب أن تتضمن الشروط التالية:

(١) السعي قدر الإمكان إلى تدوين وتوثيق حدود الترميز وأنظمته التي تظهر في تحليل اللغة؛ على سبيل المثال ما قام به كل من جوهاسون (Johansson:1982)، وسامبسون (Sampson:1987) من وضع معايير قياسية لأنظمة التحليل اللغوي، وتحوي اللغة الصينية نظامًا لترميز أنواع الكلمات وتقسيمها داخل النصوص الصينية[١].

(٢) ضرورة التوافق بين أنظمة التحليل المختلفة، بمعنى محاولة استخدام علامات معروفة لدى الجميع، وتقديم المعلومات اللازمة للتحويل من نظام ترميز إلى نظام آخر.

(٣) لا يوجد توازن بين مستويات تطور أنظمة الترميز على مستويات اللغة المختلفة. فالمستويات الأسرع تطورًا هي عمليات الترميز على مستوى المفردات ومستوى النحو ومستوى الصوتيات ومخارج الأصوات وغير ذلك. وينبغي الاهتمام بصورة أكبر من الآن فصاعدًا بأنظمة الترميز على المستويين الدلالي والوظيفي.

٣- تطور أدوات معالجة المادة اللغوية

يعتمد تحليل المادة اللغوية على بنية الحاسب بصورة أساسية، ويُقصَد ببنية الحاسب تلك الأدوات والبرمجيات التي تقوم بالبحث ومعالجة المادة

---

(١) نظرًا إلى طبيعة اللغة الصينية التي تتكون كلماتها من رموز وليس من حروف كاللغات الأبجدية يحتاج الحاسب دائمًا إلي وسائل برمجية لتمييز حدود الكلمات وفصلها عن بعضها في النصوص المكتوبة باللغة الصينية. (المترجم)

اللغوية وتطوير استخدام أساليب علم الإحصاء في بناء أدوات فعالـــة لعمـــل المعالجة الآلية للغة، من أجل زيادة فاعلية الباحثين وإنتاجيتهم فـــي مجـــال البحث اللغوي. وعلى الرغم من أن برمجيات المعالجة المتاحة حاليًا قليلـــة العدد، فإن معظمها تمت صياغته لمعالجة ذخائر لغويـــة محـــددة، ونطـــاق استغلالها محدود للغاية ولا يمكن تعميم استخدامها.

الفصل الرابع

دور الحاسب في علم الذخائر اللغوية

إذا تم التحليل اللغوي على نصوص كبيرة الحجم بصورة يدوية فمن السهل حدوث أخطاء، ومن الصعب الانتهاء من عملية التحليل، بالإضافة إلى زيادة إمكانية التكرار وغير ذلك من الصعوبات. وعلى الرغم من أن هذه الطريقة في البحث اللغوي قد قدمت إسهامات عظيمة للدرس اللغوي على مدى عدة قرون مضت وخاصة في مجال تأليف المعاجم، فإن ظهور الحاسب الآلي مع منتصف القرن العشرين قد أدى إلى حدوث تَغَيُّر جذري في مجال الدراسات اللغوية القائمة على النصوص اللغوية. ودفعتنا ثورة المعلومات إلى تغيير أساليب العمل القائمة على المواد اللغوية بصورة جذرية. ولم تعد هناك حاجة إلى استخدام البطاقات في جمع المواد اللغوية وتصنيفها؛ حيث أصبح مصممو المعاجم وعلماء اللغة يلجئون إلى استخدام الحاسب في تخزين كميات هائلة من النصوص، ويمكنهم أيضًا البحث عن المفردات أو العبارات أو الجمل وإظهارها على شاشة الحاسب بسرعة دون أدنى خطأ. والأكثر من ذلك أنه يمكن إعادة التصنيف والترتيب بناءً على أكثر من أسلوب مثل تصنيف الكلمات طبقًا لأساليب الاقتران فيما بينها أو الطبيعة النحوية لكل منها. ولذلك أصبح هناك ارتباط وثيق بين الحاسب وعلم الذخائر اللغوية. والحقيقة التي لا مراء فيها هي أن استخدام الحاسب بالفعل قد جعل البحث

اللغوي القائم على الذخائر اللغوية يتسم بالسرعة والدقة والقدرة على معالجة نصوص ذات أحجام هائلة. إن استخدام برمجيات الحاسب لم تقلل فقط من العبء اليدوي الذي كان يبذله علماء اللغة في تأليف المعاجم ومعالجة الكميات الكبيرة من المواد اللغوية؛ ولكن تلك البرمجيات قللت أيضًا من العوامل البشرية التي تؤدي إلى الخطأ في أثناء العمل. وبالإضافة إلى إمكانات الحاسب التي سبق الإشارة إليها من حيث القدرة على تصنيف المواد اللغوية وإحصائها وترتيبها، يستطيع الحاسب أيضًا أن يقدم إحصاءات سليمة عن معدل تكرار تلك المواد داخل النصوص، بالإضافة إلى أن تلك الإحصاءات الرقمية تمكننا من تكوين معلومات قياسية عن اللغة ثم تطبيقها واستخدامها في عمل برمجيات كتابة الرموز الصينية باستخدام الحاسب[1] أو تلك التي استخدمت في تمييز الأصوات اللغوية أو فهرسة النصوص، وغير ذلك من البرامج التطبيقية التي اعتمدت على المعلومات اللغوية كأساس لها.

كما يُستَخدَم الحاسب الآلي في تعظيم قدرات علماء اللغة في البحث اللغوي، وبذلك يتمكن هؤلاء العلماء من تنويع إمكانات استخدام اللغة وتدعيمها في عمل تطبيقات لغوية متقدمة. إن الدراسات الكمية المعتمدة على الذخائر اللغوية في حقل اللسانيات تساهم في زيادة الوصف العلمي للظواهر

---

(١) اعتمادًا على المعلومات الإحصائية التي تمت على ذخائر لغوية للتعرف على معدلات تواتر سلاسل الرموز والكلمات داخل النصوص الصينية تمكن العلماء الصينيون من عمل برمجيات لكتابة اللغة الصينية على الحاسب تقوم بإدخال جمل كاملة إلى الحاسب دفعة واحدة عن طريق كتابة الأحرف الأولى لنطق الكلمات داخل تلك الجمل، وبذلك احتلت اللغة الصينية مركز الصدارة بين لغات العالم من حيث كونها أسرع لغة تكتب على الحاسب. (المترجم)

اللغوية وتعميقه، بالإضافة إلى زيادة الارتباط بين ذلك الوصف وعمل التطبيقات اللغوية المختلفة في جميع المجالات. ومن بين تلك التطبيقات اللغوية الترجمة الآلية، والتحويل بين الأشكال النصية المختلفة (مثل تحويل النصوص المكتوبة إلى مكافئها الصوتي وهي تلك التقنية المعروفة اختصارًا باسم TTS (Text To Sound)، وتحليل المحتوى، وتعليم اللغات وغير ذلك من المجالات التي تُعْتَبَر المستفيد الأول من تطبيقات التحليل الكمي للذخائر النصية.

إن تحقق الحلم بمعالجة النصوص اللغوية باستخدام الحاسب إنما يرجع إلى سبعينيات القرن العشرين، وقد نشر العالم جوفندانكوتي في مجلة "الدراسات الحاسوبية للأدب واللغة" (Literary and Linguistic Computing) مقالاً يتوقع فيه أوان عصر معالجة اللغات باستخدام الحاسب بدلاً من الممارسة اليدوية للإنسان. وقد قضى هذا العالِم هو وتلاميذُه ست سنوات من العمل الشاق تمكّن بعدها من تقديم معالجة وإدارة معلوماتية لنص لغوي يصل إلى ثلاثمائة ألف كلمة باستخدام الحاسب الآلي (Govindankutty:1973). ويرى علماء لغة الذخائر اللغوية في عصرنا الحالي أن استخدام الطرق اليدوية في تدوين النصوص على الورق أصبح من الطرق العقيمة، وأن تلك الأعمال الشاقة يمكن الاستغناء عنها باستخدام لوحة المفاتيح التي يتم إدخال المعلومات من خلالها إلى الحاسب الآلي.

في منتصف ثمانينيات القرن العشرين أصبح بمقدور علماء الذخائر اللغوية إنهاء العمل في إدارة المعلومات اللغوية داخل حاسبات كبيرة الحجم

بسرعة نسبية، وقد كان مثل ذلك العمل يحتاج وقتًا طويلا في السبعينيات؛ حيث كان الحاسب يستغرق ما يزيد عن ساعة للبحث عن معدل تكرار كلمة مثل "When" في ذخيرة لغوية يصل عدد كلماتها إلى مليون كلمة، أما مع حاسبات الثمانينيات فقد أصبح هذا العمل لا يستغرق أكثر من عدة دقائق. أما الآن فبإمكان الحاسبات الشخصية أن تحتوي أقراصًا صُلْبة ذات سعة تخزينية كبيرة ووحدات معالجة ذات سرعة هائلة ويمكنها القيام بأعمال تخزين وتحليل للنصوص بصورة أكثر سهولة وملائمة.

وقد سبق أن ساهم خبراء الحاسب في بداية ثمانينيات القرن العشرين في حل بعض الإشكاليات الفنية التي كانت تعترض علماء الذخائر اللغوية في مجال استخدام الحاسب في التحليل اللغوي. ومع بداية التسعينيات ساهم التطور السريع للحاسبات والبرمجيات المتخصصة في تحليل الذخائر اللغوية وتسويقها بصورة تجارية في إزالة الخوف أمام العلماء وتشجيعهم على استخدام الحاسب وابتكار البرمجيات المساعدة على تحليل الذخائر اللغوية.

الفصل الخامس

مجالات البحث في علم الذخائر اللغوية

المستهدف الأساسي من علم الذخائر اللغوية هو النصوص اللغوية، وهذه النصوص هي المصدر الرئيس في تقديم الوصف والبراهين اللغوية. ومن بين أنواع ذلك الوصف، تقديم الوصف الكمي لتوزيع المفردات والتعبيرات داخل اللغة. وهذا المنحى البحثي تَطَوَّر تدريجيًا ليصبح أحد المكونات الأساسية التي لا غنى عنها في الدراسات اللغوية، وهذا يتشابه تمامًا مع ما أشار إليه ليتش **Leech** عام ١٩٩٢ قائلاً: "إن الهدف من البحث اللغوي هو وصف الاستخدام اللغوي وليس وصف القدرة اللغوية؛ لأن متابعة الطبيعة اللغوية ومراقبتها وسط بيئة الاستخدام هي التي تؤدي إلى ظهور النظرية اللغوية وليس العكس".

ومن ثَمَّ نجد أن علم اللغة يختلف عن النظرية اللغوية عند التحويليين التوليديين فهذا العلم ليس نظرية لغوية، ولا يمكن أن نطلق عليه أنه فرع مستقل أو جديد من أفرع علم اللغة، وإنما هو الوسيلة المنهجية التي يتبعها علماء اللغة في إجراء البحوث اللغوية. إن علماء اللغة في أثناء بحثهم في جوهر اللغة وعناصرها وأبنيتها ووظيفتها يحتاجون إلى أدلة وقرائن لغوية لوصف ما هو ممكن داخل اللغة. وهذه القرائن اللغوية يستخلصها علماء اللغة من مراقبتهم ودرسهم للظواهر اللغوية في عصور مختلفة، ويتأتى ذلك

إما عن طريق الاستقراء والجمع وإما عن طريق مراقبة اللغة ووصفها في مستوييها الشفهي والتحريري. وفي الدراسات القائمة على الذخائر اللغوية يمكن استخلاص تلك الحقائق بصورة مباشرة من النصوص. وفي تلك النقطة نجد أن علم الذخائر اللغوية يختلف عن النظرية اللغوية التي تعتمد على القرائن الجدلية والمنطقية التي يتبعها علماء اللغة التقليديُّون، فعلم الذخائر اللغوية لا يقدم احتمالات ظهور المفردات والأبنية والوظائف المختلفة داخل اللغة فحسب، بل يقدم أيضًا معدلات ظهور تلك الصور اللغوية داخل اللغة. وهذا العلم يتشابه مع باقي علوم اللغة في كونه يبحث في طبيعة اللغة وأبنيتها ووظيفتها كما يبحث أيضًا في وسائل اكتساب اللغة والتحول اللغوي وتَغَيُّر اللغة. ونقاط التركيز في هذا العلم هي المفردات والوظائف النحوية لتلك المفردات وليس النحو بمعناه الضيق.

١- إنشاء الذخائر اللغوية وتصميمها

الذخائر اللغوية هي المصدر الذي يمد علم الذخائر اللغوية بالمادة اللغوية؛ ولذلك فإن تصميم الذخائر اللغوية وتأليفها هو أساس البحث في هذا العلم. وهذا التصميم يضم في معناه تصميم الذخيرة اللغوية ذاتها، ووسائل جمع المادة اللغوية وتسجيلها وإدارتها. ولا يُقصد بالذخيرة اللغوية ذلك التجميع السطحي للنصوص، وإنما يقصد به أن تكون تلك النصوص مُعَبِّرَة عن اللغة بصورة عامة، أو في أحد التخصصات. ولذلك فعند تصميم ذخيرة لغوية ينبغي أن تكون البداية بالعمل طبقًا لهدف عام يحكم بناء المادة اللغوية والتفكير بصورة دقيقة في المبادئ التي ستُتَّبَع في اختيار العينات اللغوية

وتوزيع تلك العينات داخل إطار اللغة؛ وذلك لنتمكن قدر المستطاع من جعل تلك الذخيرة اللغوية تمثل اللغة في فترة زمنية محددة.

ونظرًا إلى عدم وجود مؤشرات كمية ثابتة توضح لنا الشروط التي يمكن معها اعتبار العينات اللغوية ممثلة للغة في إطارها الأشمل، فليس أمامنا إلا أن نستعين بتجاربنا الذاتية وخبرات من سبقونا في بناء الذخائر اللغوية. وانطلاقًا من هذا المفهوم، يمكننا أن نقرر أن قواعد اختيار العينات اللغوية ونسب توزيعها داخل إطار اللغة من الموضوعات التي يتوقف عليها تطور الذخائر اللغوية التي لم يتم التوصل إلى حلول وإجابات لها حتى الآن. وبطبيعة الحال، فإن نطاق الذخائر اللغوية ومقاييس حفظها وتخزينها وما إلى ذلك تعد قضايا مطروحة ويجب الإجابة عليها في أثناء عملية إنشاء الذخائر اللغوية. ومن دون ذلك لا يمكن أن نجعل من الذخائر اللغوية مصادر حقيقية يمكن الاعتماد عليها في دراسة اللغة بصورة شاملة ومتكاملة.

٢ــ تقنيات إدارة الذخائر اللغوية ومعالجتها

يقصد بذلك تلك الأدوات البرمجية التي تُستَخدم في تحليل المادة اللغوية وترميزها، وحمايتها، وفهرستها. فالذخيرة النصية لا تقتصر فقط على كونها وعاءً يحتوي نصوصًا لغوية، وإنما ينبغي أن تتمتع بقدرات متميزة في التخزين والاسترجاع، حتى يمكن أن تتيح لجميع الباحثين فرصة الحصول على المعلومات التي يرغبون فيها من المادة اللغوية المخزنة داخل الذخيرة اللغوية؛ ولذلك فإن فهرسة المواد اللغوية من الأعمال المهمة. ومن أشهر التقنيات المستخدمة في الفهرسة تلك التقنية المعروفة باسم **Concordance**.

لقد حدث تَحَوّلٌ هائلٌ في نطاق الذخائر اللغوية وأنواعها وسهولة التعامل معها وغير ذلك طبقًا لزيادة فاعلية أداء الحاسبات. وبالفعل، حدث تطوير لعدد كبير من البرمجيات التي تقوم بمعالجة البيانات وتداولها داخل الذخائر اللغوية. إلا أن قيمة الذخيرة اللغوية باعتبارها مصدرًا للمادة اللغوية موضع البحث لا يمكن أن تُقاس فقط من حيث الحجم، وإنما ينبغي أن يكون عنصر الترميز فيها هو المصدر الذي يحمل قيمة مضافة إلى الذخائر. بمعنى أن يتم ترميز الذخائر اللغوية بأساليب مختلفة، وعن طريق تلك الوسائل الترميزية يمكن تعظيم نطاق المعلومات وتوسيعه داخل الذخيرة بصورة واضحة. ومن ثَمَّ، نقدم إنجازات أكثر قيمة للباحثين في مجالات اللغة كافة. وبالنسبة إلى الذخائر اللغوية الصينية فإن عملية تمييز حدود الكلمات[١] تعتبر بخلاف باقي اللغات الهندأوروبية هي الخطوة الأولى التي تُبْنَى عليها جميع مستويات المعالجة اللغوية، ثم تأتي بعد ذلك مستويات المعالجة الأخرى التي تتم مع تلك اللغات؛ مثل: الترميز على مستوى نوع الكلمة، والترميز على مستوى دلالة الكلمات ودلالة الجمل، والترميز على مستوى النحو، بالإضافة إلى الترميز على مستوى الفقرات والنصوص الكاملة، وغير ذلك من عمليات المعالجة.

---

(١) بمعنى إضافة مسافة تفصل بين الكلمات وبعضها داخل النصوص الصينية، نظرًا إلى طبيعة اللغة الصينية التي تتكون من رموز متراصة مع بعضها دون مسافات، ويعتمد القارئ على خبرته اللغوية في وضع تلك الرموز معًا في إطار كلمات، أما في حالة التعامل الحاسوبي مع اللغة الصينية فتكون هناك حاجة ماسة لوجود تلك المسافات. (المترجم)

٣- استخدام الذخائر اللغوية في البحث اللغوي

"إن ظهور العديد من وسائل تطوير معالجة المادة اللغوية واستخراج المعلومات اللغوية من بين النصوص قد زاد وعمَّق قدرتنا على معرفة اللغة وفهمها". هاليداي (Halliday:1991)[12] لقد قَدَّمَت الذخائر اللغوية مصادر معلومات قيمة مَكَّنَت العلماء من وصف اللغة، وقد استطاع علماء اللغة تقديم وصف للمفردات والقواعد داخل اللغة اعتمادًا على تلك الذخائر اللغوية المُخَزَّنة على وسائط إلكترونية؛ وذلك استنادًا إلى فرع اللغة الذي يتخذ من الذخائر اللغوية نقطة ارتكاز له. ولم تقتصر موضوعات الدراسة على مراقبة إمكانات ظهور الكلمات داخل النصوص، بل تطرقت إلى دراسة حالات ظهور تلك الكلمات أيضًا. وبالنسبة إلى الدراسات المتعلقة بتوزيع المفردات والقواعد داخل النصوص، فقد ساهمت الذخائر اللغوية في دفع بحوث تصنيف النصوص، والتحوُّل اللغوي وتلك الدراسات المتعلقة بتغير أشكال اللغة. وعن طريق استخلاص المعلومات من مواد لغوية كبيرة الحجم، قدمت الذخائر اللغوية معلومات وافرة عن السياقات النصية من شأنها أن تخدم مجال الدراسات الدلالية للغة.

٤- تطبيقات الذخائر اللغوية في علم اللغة الحاسوبي

إن التطبيقات التي يقدمها الوصف اللغوي القائم على الذخائر اللغوية هي ذلك المجال المليء بالإبداع العلمي والتطبيقي في علم الذخائر اللغوية. فمن الممكن استخدام نتائج البحث في علم الذخائر اللغوية لعمل تطبيقات في

معالجة اللغات الطبيعية وتمييز الأصوات اللغوية ونُظُم الترجمة الآلية. ومع بداية التسعينيات، شغلت الدراسات الإحصائية التي تمت على ذخائر نـــصية كبيرة الحجم مركز الصدارة تدريجيًّا في مجالات معالجة اللغات الطبيعيـــة. مجمل القول: إن فريقًا من العلماء في علم الذخائر اللغوية يتولى البحث فـــي أساليب تصميم الذخائر اللغوية ومنهجياتها، وهناك فريق آخر مـــنهم مهـــتم بالبحث في وسائل تحليل النصوص ومنهجيات معالجتها، وهناك فريق ثالث وهو القسم الأكبر الذي يهتم بالبحث في توصيف اللغة داخل الذخائر اللغوية وبحث تطوير التطبيقات القائمة على ذلك.

الفصل السادس

حول إعداد هذا الكتاب

الباب الأول: من هذا الكتاب يُعَرّف القارئ بالمفاهيم الأساسية حول علم الذخائر اللغوية، ويستعرض بصورة مُبَسَّطة تاريخ تطور هذا العلم، بالإضافة إلى شرح لمجالات البحث في علم الذخائر اللغوية. أما الأبواب التالية من الكتاب فتتناول تلك المجالات بمزيد من الشرح والتحليل.

الباب الثاني: يتناول بالتفصيل بناء الذخائر اللغوية ويتعرض بمزيد من التعمق لبعض القضايا التي يجب أخذها بعين الاعتبار عند العمل في الذخائر اللغوية؛ بغرض إقرار بعض المبادئ الواجب توافرها عند بناء الأنواع المختلفة للذخائر اللغوية، وفي نهاية الباب نُعَرّف بعض الذخائر اللغوية المهمة داخل الصين وخارجها.

أما **الباب الثالث:** فيتعرض للتقنيات الخاصة بإدارة الذخائر اللغوية ومعالجتها في إطار ثلاثة محاور أساسية: تقنيات فهرسة الذخائر اللغوية (فهرسة الكلمات)، وأساليب الإحصاء المستخدمة مع الذخائر اللغوية، وترميز الذخائر وما إلى ذلك. وتعتبر النقطة الثالثة هي المحور الرئيس لهذا الباب.

أما **الباب الرابع:** فيتناول الدراسات اللغوية القائمة على الذخائر اللغوية ويتعرض إلى وسائل الوصف الكمي المستخدمة في وصف المستويات

المختلفة للغة بغرض التعبير عن خصائص اللغة عند الاستخدام. وفي النهاية، يقدم هذا الباب بعض الأمثلة الحقيقية عن تطبيقات الذخائر اللغوية على اللغتين الصينية والإنجليزية.

ويركز الباب الخامس: على تطبيقات علم الذخائر اللغوية في مجال علم اللغة الحاسوبي، بمعنى كيفية الاستعانة بنتائج البحث في علم الذخائر اللغوية لعمل تطبيقات تخدم مجالات معالجة اللغات الطبيعية كافة.

الباب الثاني

تصميم الذخائر اللغوية وتطويرها

إن أول مَهَمَّة يُكَلَّف بها من يعمل في حقل علم الذخائر اللغويـــة هـــي عملية بناء الذخيرة، ومن هنا ينبغي له أن يَتَّخذ مجموعـــة مـــن القـــرارات الحاسمة بشأن المادة اللغوية التي ستتضمنها الذخيرة اللغوية ووسائل جمـــع تلك المادة وغير ذلك من الإجراءات، هذا بالإضافة إلى ضرورة أن يتمتـــع من يقوم ببناء الذخيرة على القدرة على التحكم في المشاكل التي قد تظهر في مراحل إدارة الذخيرة اللغوية بعد تأسيسها. أما علماء اللغة الذين يتعاملون مع الذخيرة بعد بنائها فعليهم أن يتمتعوا بالقدرة على معالجة جميع الأمثلة اللغوية التي تظهر في الذخيرة اللغوية.

الفصل الأول

تصميم الذخائر اللغوية وإشكاليات تجميع المادة اللغوية

إن الغرض الأساسي الذي يسعى إليه علماء اللغة من إنشاء الذخائر اللغوية هو استخدام تلك الذخائر في عمليات استقراء التراكيب اللغوية وتحليلها والتعرف على طرق استخدامها، فقد انتبه جوهانسون (١٩٩٤)[14] في منتصف تسعينيات القرن العشرين إلى أن أكثر فعل يقترن مع كلمة "Corpus" هو فعل "Compile" بمعنى تجميع، وبالفعل فقد كانت عملية تجميع المادة اللغوية داخل الذخائر اللغوية ووضع هيكلية بنائها وتحديد نطاق الذخائر اللغوية من الموضوعات التي ظلت تشغل علماء الذخائر اللغوية طوال الفترة من ستينيات حتى سبيعنيات القرن العشرين.

إن نقطة الانطلاق التي تبدأ منها عملية تصميم الذخائر اللغوية وتجميعها هي: كيفية جعل المعلومات اللغوية التي ستُستخرج من تلك الذخائر منطقية ويمكن الاعتماد عليها. ولذلك أشار كينيدي (Kennedy:1998)[15] إلى أول مشكلة تواجه مصممي الذخائر اللغوية، ألا وهي إلى أي مدى تُعْتَبَر البيانات اللغوية التي تحتويها الذخيرة مُعَبِّرَة حقًا عن البناء اللغوي المطلوب دراسته. إن هذه المشكلة يمكن النظر إليها من خلال عدة زوايا؛ منها: هل العينات اللغوية للذخيرة ستكون عينات ثابتة أم متغيرة (استاتيكية أم

ديناميكية)؟ وإلى أي مدى يمكن اعتبار حجمها مُمَثِّلا لبنية اللغة؟ وما الحجم الذي ينبغي أن تكون عليه الذخيرة حتى يمكن أن تفي بالاحتياجات العامة أو المتخصصة في أغراض البحث اللغوي؟ وكم عدد العينات التـي تحتويهـا الذخيرة اللغوية؟ وما حجم كل عينة؟ وفيما يلي نقدم مناقشة تفـصيلية لكـل سؤال من تلك الأسئلة التي طرحها كينيدي.

أولاً: الاستاتيكية والديناميكية

الذخيرة اللغوية يمكن جمعها بالاعتماد على نصوص استاتيكية ثابتـة، والغرض من ذلك محاولة الحصول على عينات لغوية تكون ممثلة للغة فـي فترة زمنية محددة. والذخيرة اللغوية التي يصل حجم مادتها مليـون كلمـة تندرج تحت هذا النوع من الذخائر اللغوية. مثال على ذلك ذخيرة إس إي يو SEU التي حاولت اختيار عينات للنموذج البريطاني من اللغة الإنجليزية في ظروف استخدام مختلفة على المستويين التحريري والشفهي بصورة ثابتة مما يجعل من تلك الذخيرة اللغوية مصدرًا مُمَثِّلا للغة الإنجليزية بصورة عامـة. وعند تصميم هذا النوع من الذخائر اللغوية دائمًا ما نكون حريـصين عنـد التعامل مع بعض القضايا، مثل أسلوب الكتابة ونطاق جمع العينات وغيـر ذلك. وتُعتَبر الدراسة التي أصدرها كـويرك (Quirk) وآخـرون[16] عـام ١٩٨٥ التي سُمِّيَت باسم "موسوعة قواعد اللغة الإنجليزية" (Comprehensive Grammar Of English) ، قد تمت كتابتها بناءً على ما قدمته ذخيـرة إس إي يو (SEU) من معلومات لغوية، وكان كويرك يعتقـد أن هـذه الـذخيرة

اللغوية أشبه بمن يلتقط صورة سريعة للإنجليزية البريطانية، فهي مثل صورة لمنظر طبيعي تم التقاطها بسرعة، بغرض تسجيل الملامح الرئيسة لذلك المشهد. وعلى الرغم من أن مصممي تلك الذخيرة استخدموا أسلوب العينات كنمط رئيس لبنائها، فإنهم لم يستطيعوا حصر جميع الظواهر اللغوية في متنها. وفي الحقيقة فقد جَمَعَت بعض النصوص العامة؛ ولذلك فهي لم تأخذ عن اللغة إلا سلسلة من اللقطات والشرائح اللغوية بشكل استاتيكي. والحقيقة أن هذا النوع من الذخائر اللغوية أشبه بمحاولة تجميد مجرى تطور اللغة في لحظة بعينها. ولكن نظرًا إلى أن مصممي الذخائر دائمًا ما يستعينون بعينات لغوية ونماذج نصية محددة العدد في بناء الذخائر اللغوية، فمن الممكن أن يتم مقارنة ذلك النوع من الذخائر القائمة على أسلوب البناء عن طريق العينات مع نماذج أخرى من الذخائر اللغوية مشابهة من حيث طريقة البناء. ومن الممكن النظر إلى كل من الذخائر اللغوية صغيرة الحجم أو الكبيرة على أنها ذخائر لغوية استاتيكية على حدٍّ سواء، لدرجة أننا نستطيع القول: إن ذخيرة بي إن سي الإنجليزية **BNC** كبيرة الحجم التي يصل عدد كلماتها إلى مائة مليون كلمة من النصوص اللغوية الاستاتيكية.

الفكرة الأخرى التي تُطرَح عند بناء الذخائر اللغوية هي تلك الفكرة القائمة على بناء ذخائر لغوية ديناميكية متغيرة أو ذخائر لغوية لمراقبة التغيرات اللغوية **(monitor corpus)**[17]، وهذا النوع من الذخائر أشبه بالصور المتحركة ولا تُعتَبَر لقطة سريعة تؤخذ للغة في لحظة ما. وقد أُطلِقَت تلك التسمية على هذا النوع من الذخائر لأنها تُقدِّم وسيلة فعالة تمكننا

من مراقبة نماذج الاستخدام اللغوي مقترنًا بحالات التغير التي قد تلحق باللغة مع مرور الوقت. وهذا النوع من الذخائر يَعْتَمِد على تجميع كميات هائلة من النصوص اللغوية في أوقات معينة. ثم الاستعانة ببعض البرمجيات للبحث عن معلومات إحصائية طبقًا لأهداف وصفية معينة يرغب الباحثون في التوصل إليها، هذا بالإضافة إلى إمكانية استخلاص بعض النتائج عن تلك الظواهر اللغوية التي يتم التوصل إليها واستنتاجها، على سبيل المثال ملاحظة تواتر بعض التراكيب الجديدة وأنماط الكلمات داخل اللغة، أو حدوث بعض التغيرات في استخدام الكلمات القديمة، أو في قدرة تلك الكلمات على الاقتران مع كلمات أخرى. وهكذا يقدم النمط الديناميكي للذخائر اللغوية معلومات متلاحقة عن التغير الذي يطرأ على اللغة في النصوص الحديثة. وقد وصف سنكلير (Sinclair) هذا المفهوم عن الذخائر اللغوية الديناميكية قائلاً: "إننا كما لو كنا نُسْبِر أغوارَ التحول اللغوي الذي يحدث مع سريان اللغة في مجرى الزمن". إلا أن الذخيرة اللغوية التي يصل عدد مفرداتها إلى مائة مليون كلمة تبدو كبيرة جدًّا على خطوات معالجة أية ظاهرة لغوية. ونظرًا إلى أن العناصر التي تتكون منها الذخيرة اللغوية، وطرق الاستخدام اللغوي لتلك العناصر دائمًا ما تتغير بشكل مستمر، فإن هذا يجعلنا غير قادرين على إجراء دراسات مقارنة بين الذخائر اللغوية عندما يتغير نوع النصوص المحتواة في داخلها، مثل مقارنة التحول في معدل تكرار الكلمات عندما يتغير نوع النص. هذا بالإضافة إلى أن المادة اللغوية في الذخيرة الديناميكية يتم جمعها بصورة عفوية ولا يكون هناك ضمان لتوازن تمثيل العينات في كل التخصصات.

ففي تلك الأحوال، يحل الاهتمام بكمية المادة اللغوية وحجمها تلقائيًّا محل التصميم الدقيق والتخطيط المحكم لطريقة جمع العينات اللغوية. هذا بالإضافة إلى جمع تلك النصوص ومعالجتها وتخزينها على وسائط التخزين الصلبة داخل هيكلية الحاسب. الأمر الذي يحتاج إلى تكاليف مادية عالية تتمثل في الحاجة إلى خبرات تقنية وبرمجيات معقدة لإجراء عمليات التحليل والمعالجة؛ ولذلك فإن فرص ممارسة الأبحاث العلمية القائمة على استقراء الذخائر اللغوية وتحليلها تُعدُّ قليلة أمام جمهور العلماء المتخصصين منهم والعامة. ولا يقدر على مثل هذا النوع من الذخائر اللغوية إلا المؤسسات العملاقة والهيئات الحكومية وبعض الجهات المعنية المتخصصة. ويمكن لبعض الباحثين المنفردين التعامل مع تلك الذخائر اللغوية نظير مبالغ معينة تُدْفَع كرسوم حتى يُسْمَح بالتعامل مع المادة اللغوية المدونة بداخلها، ونجد أن غالبية البحوث القائمة على ذخائر لغوية يقوم بها أصحابها بصورة منفردة مثلهم مثل معظم القائمين بأبحاث لغوية تقليدية، وقليلا ما يكون هناك تعاون بين مجموعات عمل. ولا يتم ذلك إلا عن طريق تبادل الأفكار بين علماء اللغة فيما يخص تقنيات الحاسب الآلي وتبادل النصوص. ولذلك إذا أردنا للذخيرة الديناميكية أن يقتصر استغلالها على قلة من العلماء فسوف يكون ذلك بمثابة إهدار كبير للمال والجهد.

إن اللجوء إلى استخدام الذخائر اللغوية الديناميكية يُقَدِّم لعلماء اللغة رؤية دقيقة وواقعية عن اللغة (على الأخص في مجال صناعة المعاجم وأبحاث علم اللغة التاريخي)، وتتمثل تلك الرؤية في رصد مراحل التغير

اللغوي ووسائل استخدام الكلمات التي لها معدل استخدام منخفض. ولكن أيًا ما كان الأمر فإن هذا النوع من الذخائر اللغوية قادر أيضًا على ضمان إمكانية دراسة الكلمات ذات معدل التكرار الأعلى والمتوسط أيضًا، فضلاً عن البحوث اللغوية المتعلقة بالصوتيات والصرف والنحو.

وفي الوقت الحالي أدى النمو المضطرد في قدرة الحاسب على تخزين كم هائل من البيانات ومعالجتها إلى إتاحة التعامل مع المواد اللغوية فائقة الحجم التي تتعدى مائة مليون كلمة. بالإضافة إلى أن هناك حاجة ماسة لوجود ذخائر لغوية بمثل هذا الحجم لاستخدامها في التطبيقات المتعلقة بتمييز الأصوات وتحويل الأصوات اللغوية إلى مكافئها التحريري وغير ذلك. ولهذا نعتقد أن بناء الذخائر اللغوية حاليًا ينبغي أن يقترن بسعي حثيث نحو جمع ما يمكن جمعه من المواد اللغوية حتى نتمكن من بناء ذخائر لغوية فائقة الحجم. إلا أن مشكلة إنشاء هذا النوع من الذخائر اللغوية تتمثل في كيفية التوصل إلى وسيلة لتنظيم المواد اللغوية وتصنيفها في تلك الذخيرة حتى يمكن إتاحة الفرصة لإعادة استدعائها بما يتناسب مع التوجهات المختلفة للبحث اللغوي، ومساعدة الباحثين على تكوين ذخائر نصية متنوعة من رحم الذخيرة الأم لأغراض بحثية معينة. على سبيل المثال، إذا أردنا أن نستطلع قوائم المفردات والمصطلحات المستخدمة في مجال ما، يمكننا أن نجتزئ من الذخيرة الأم ذخيرة متخصصة تساعدنا على استخراج تلك المعلومات.

ثانيًا: المرجعية والتوازن

الموضوع الآخر المتصل باستاتيكية الذخيرة اللغوية أو ديناميكيتها هو الشروط الواجب توافرها في النصوص اللغوية حتى يمكن اعتبارها مناسبة

يمكن الاعتماد عليها في البحث اللغوي؛ وفي الحقيقة يُعْتَبَر هذا التساؤل من الموضوعات الشائكة، حتى الآن لم يتوصل العلماء إلى رأي موحَّد حول المقاييس التي تُغْطِي للذخيرة اللغوية مرجعية علمية ومعايير مُحَدَّدة لتحقيق توازن النصوص في داخلها.

وقد سبق أن أشار ليتش (Leech: 1991) [18] إلى أن مدى مرجعية الذخيرة اللغوية يعود إلى قدرتها على استخلاص معلومات لغوية تمثل اللغة في إطارها العام أو المتخصص حسب توجهات البحث اللغوي. وقديمًا، كانت الهيكلية التي بُنيت على أساسها ذخيرة براون (Brown) أو ذخيرة لوب (LOB) تعتمد على مراعاة الدقة في التصميم؛ ولذلك تم الاتفاق على أن المادة اللغوية في هاتين الذخيرتين تمثل مرجعية للنموذجين البريطاني والأميركي من اللغة الإنجليزية كلٍّ على حدة.

وفي حقيقة الأمر، توصل العلماء مؤخرًا إلى حكم على مقاييس المرجعية ومعايير التوازن في الذخائر اللغوية، إلا أن ذلك الحكم لا يتعدى كونه حكمًا تقريبيًّا وليس قاطعًا. أما جوهر المشكلة فهو: ما الشيء الذي تمثله الذخيرة اللغوية؟ لقد بحث العلماء عشرات السنوات في تحليل الكلام والدراسات المتعلقة بعلم اللغة الاجتماعي، وعلى الرغم من أن العينات اللغوية لا تكفي للتعبير عن أسلوب أو موضوع لغوي معين، فإنه ما زال هناك من يعتقد أن الذخيرة اللغوية المكوَّنة من كم كبير من النصوص قادرة على أداء تلك المهمة. خلاصة القول: إن تلك النماذج اللغوية يمكن استخدامها فقط للتعبير عن جوهر اللغة. على سبيل المثال عندما نتناول علم الصوتيات

في لغةٍ ما، على الرغم من عدم إمكانية تشابه الأصوات بين متحدثي اللغة الواحدة، فإننا نستطيع استخلاص النظام الصوتي لتلك اللغة دون أدنى صعوبة. وعلى هذا فإن بناء المعاجم الكبرى والوصف النحوي لهيكل لغة ما لا يخرج عن هذا المفهوم الذي يقدم وصفًا إجماليًا للغة.

وهناك إشكالية أخرى تظهر عند تصميم الذخائر اللغوية وهي كيفية ضمان تحقيق توازن بين الموضوعات المختلفة المكوّنة لبنية الذخيرة اللغوية؛ فقد كانت غالبية الذخائر اللغوية في المراحل الأولى تنحاز للمستوى التحريري من اللغة، وتُعطي السلطة الأعلى لهذا الشكل اللغوي من النصوص، لدرجة أنهم كانوا لا يستخدمون إلا النصوص التحريرية لبناء الذخائر اللغوية. والسبب في ذلك يرجع إلى سهولة التعامل مع النصوص المكتوبة باستخدام الوسائط الإلكترونية(١) لدرجة أنه في الجيل الثاني من الذخائر اللغوية كبيرة الحجم مثل ذخيرة بي إن سي BNC التي يصل حجمها إلى مائة مليون كلمة لم تتجاوز نسبة المصادر اللغوية الشفهية بها مقدار ١٠% من إجمالي حجم الذخيرة. وعلى العكس من ذلك نجد أن بعض الذخائر اللغوية مثل ذخيرة آي سي إي ICE تُمَثَّل المادة الشفهية بها نسبة ٦٠% من إجمالي حجمها، أما المادة التحريرية فتمثل ٤٠% فقط. وهذا النوع من الذخائر يمثل أقلية من بين إجمالي عدد الذخائر الموجودة حتى

---

(١) لا يخفى عن القارئ أن قدرة الحاسب الآلي على التعامل مع المستوى التحريري من اللغة قد سبقت المستويات الأخرى؛ لذلك تأخر تدوين اللغة في إطارها المنطوق داخل الذخائر اللغوية كثيرًا عن المكافئ التحريري لها. (المترجم)

الآن. حتى لو تحدثنا عن ذخيرة لغوية تحريرية؛ فمن الصعب الاستهانة بالإجابة على التساؤلات التي تتردد عن نوع المادة اللغوية التي ينبغي أن تتوفر في تلك الذخيرة. ولا توجد حتى الآن وسيلة مُتَعَارَف عليها لتصنيف أنواع النصوص المُسَجَّلَة داخل الذخائر اللغوية.

والأكثر من ذلك، فإننا لا نعرف حتى الآن ما النسبة التي تضمن لنا التوازن بين عينات النصوص في التخصصات المختلفة. لدرجة أن الذخائر اللغوية التي تم تصميمها بالفعل ليس من المنتظر اعتبارها ممَثِّلَة عن الإطار الكامل للغة، وإنما تعبر عن مجالات ونصوص وموضوعات محددة. وعلى الرغم من ذلك فإننا نجدها تعاني من مشكلة التوازن بين النصوص. وفي هذا المجال، لا يوجد إلا نوع واحد من الذخائر اللغوية لا تُثار حوله إشكالية التوازن هذه، وهو ذلك النوع الذي يتكون من المؤلفات التي نُشِرَت في فترة زمنية معينة كأن تشتمل الذخيرة على المؤلفات الكاملة لأديب معين أو نصوص كاملة لنوع معين من الفنون الأدبية. إن توازن العينات النصية داخل الذخائر اللغوية لا يمكن النظر إليه على أنه ذلك الذي يتحقق بمجرد النظر إلى مصدر المادة اللغوية كأن نهتم بتحقيق التوازن بين العينات التحريرية والعينات الشفهية؛ لأنه في حقيقة الأمر لا يوجد أحد يستطيع أن يعرف على وجه الدقة ما نسبة الكلمات الشفهية إلى الكلمات التحريرية التي تتولد في أي يوم من أيام اللغة. ولو نظرنا إلى الأمر من حيث الشكل اللغوي لوجدنا أن نسبة ما نستقبله أو نُنْتِجُه يوميًّا من الكلمات الشفهية أكبر بكثير من الكلمات التحريرية. ولكن النص المكتوب (على سبيل المثال مقال صحفي في

جريدة) من الممكن أن يقرأه ما يزيد عن عشرة ملايين قارئ، وعلى العكس من ذلك نجد أن الحوار الذي يتم بين شخص يشتري حذاء وبين البائع لا يمكن أن تتاح الفرصة لشخص آخر أن يستمع إليه، وعلى المنوال نفسه نجد أن حوارًا في نشرة الأخبار من الممكن أن يصل إلى آذان عدد أكبر من الناس مقارنة بذلك الحوار الذي يتم بين البائع ومشتري الحذاء.

إن التوازن من الأمور التي يصعب تحقيقها في الذخائر التحريرية أيضًا. وقد ذكر سِنكلير عام ١٩٩١ أن أبسط ما يجب النظر إليه من قواعد عند بناء ذخيرة لغوية تحريرية لضمان أقل حد من التوازن هو التمييز بين النصوص القصصية وغير القصصية، والتمييز بين الكتب والجرائد والمجلات، والتمييز بين المطبوعات الرسمية وغير الرسمية، بالإضافة إلى مراعاة عُمْر الكاتب وجنسه وجنسيته وغير ذلك. وهناك تساؤل آخر يتمحور حول التوازن بين الكُتَّاب الذين يمثلون الأقلية وبين القراء، من بين هؤلاء الكاتب الذي يتمتع بالسلطة الأعلى، والكاتب الذي تحظى مؤلفاته بأكبر عدد من القراء. وقد توصل مصممو الذخائر اللغوية إلى أساليب أكثر تعقيدًا وتشابكًا في محاولاتهم لتحقيق المرجعية والتوازن للذخائر اللغوية. وهكذا تُعْتَبر ذخيرة بي إن سي **BNC** خير نموذج لذلك.

ناقشت سومرز (**Summers:1991**)[19] بعض الإشكاليات المُهِمَّة التي ينبغي التفكير فيها لجعل الذخيرة اللغوية تتصف بالمرجعية، وقد تنبهت إلى أننا حتى لو أخذنا ذخيرة لغوية تحريرية يصل حجمها إلى مليون كلمة فإن تلك الذخيرة ستظل في إطار الذخائر الصغيرة جدًّا. وذلك إذا قارنّاها

بمضمون مقالات الجرائد وبالمقالات العلمية، فسوف نجد أن هناك اختلافًا بينهما من حيث طبيعة المفردات. وطبقًا لتلك الملاحظة دعت سومرز إلى البدء ببناء الذخيرة باتباع أسلوب موضوعي لاختيار العينات. ثم يُعقّب ذلك تعديل محتوى الذخيرة طبقًا لتحليل يتم عليها بغرض خلق التناسق بين المادة اللغوية لتحقيق التوازن فيما بين العينات اللغوية الواردة فيها. وقد ذكرت لنا سومرز العديد من الوسائل التي تُمَكِّننا من اختيار النصوص التحريرية، منها مدى تمتع النصوص بقيمة علمية "مدى تأثير تلك النصوص" وأسلوب اختيار العينات العشوائية، ودرجة انتشار النص وذيوعه، أو درجة إقبال القراء عليه، ومن ثَمَّ نجد أن النصوص الصحفية ستكون أكثر النصوص التي يُقْبِل عليها جمهور القراء؛ ومدى إمكانية الاطلاع على تلك النصوص، ومدى توافر معلومات إحصائية عن النصوص الأكثر تداولا بين القراء، وأخيرًا مدى الخبرة في استخدام اللغة في عمل وصف دقيق لطريقة اختيار النصوص، وما إلى ذلك. وبطبيعة الحال، فإن الممارسة العملية لبناء ذخيرة لغوية لا بد أن تجمع بين الأساليب السابقة مثل اللجوء إلى التعَرُّف على مدى تأثير قوة انتشار نوع معين من النصوص على إمكانية اختيار هذا النوع داخل الذخيرة اللغوية وما إلى ذلك.

إضافةً إلى ما سبق فإن تحديد التاريخ والفترة الزمنية التي يتم اختيار النصوص خلالها من العناصر المهمة جدًا. فالمؤلفات التاريخية الشهيرة قد لا تصبح موضع اهتمام قطاع عريض من القراء أو أنها قد تَفْقد تأثيرها. ومن ناحية أخرى نجد أن الكتب الدينية مثل الكتاب المقدس للملك جيمس (**King James**) قد تمت ترجمته منذ عدة مئات من السنين ولكنه ما زال مؤثِّرًا حتى الآن.

يلجأ مصممو الذخائر اللغوية في الصين حاليًا إلى عناوين الموضوعات والشكل الأدبي للنصوص لتحقيق التوازن داخل بنية الذخائر. وقد طَرَحَت جامعة اللغات والثقافة ببكين مؤخرًا فكرة لبناء ذخيرة لغوية كبيرة الحجم تعتمد على مفهوم درجة تداول النصوص، كمعيار لمرجعية المادة اللغوية وتوازن الموضوعات داخل الذخيرة.

ثالثًا: الحجم

كيف يمكن الوصول بالذخيرة اللغوية إلى درجة يمكن القول معها: إنها متوازنة وتتمتع بمرجعية لغوية، بالإضافة إلى إمكانية استخدامها في عمل دراسات مقارنة؟ تتوقف إجابة هذا السؤال على جودة المادة اللغوية، إلا أنه في بعض الأحيان يتم النظر إلى عدد النصوص المكونة للذخيرة، وهذا الأمر لا يتعلق فقط بعدد الكلمات داخل الذخيرة ولكن يتطرق أيضًا إلى مدى تنوع النصوص التي تحتويها الذخيرة، وعدد الكلمات التي يضمها كل نوع من هذه النصوص وما إلى ذلك. وفي الحقيقة أن نطاق المادة اللغوية ومرجعيتها يحددان مدى ملاءمة الذخيرة اللغوية للاستخدام، ومدى قدرة الباحثين على الاعتماد عليها في البحث العلمي. ولكننا يجب أن نؤكد مرة أخرى أن كِبَر حجم الذخيرة اللغوية لا يعني بالضرورة جودتها، وإنما تتحدد تلك الجودة بناءً على مدى قدرتها على أن تُمَثِّل اللغة في إطارها العام.

١- حجم المادة اللغوية

في سبعينيات القرن العشرين كانت الذخائر اللغوية التي يصل حجمها إلى مليون كلمة تبدو كبيرة الحجم، وكانت الحاسبات العملاقة في ذلك الوقت

تستغرق عدة ساعات لعمل فهرسة لكلمات إحدى الذخائر اللغوية. وفي الثمانينيات أصبحت فهرسة هذا النوع من الذخائر لا تستغرق أكثر من عشر ثوان فقط. وأصبح الجيل الجديد من الذخائر اللغوية مثل ذخيرة كوبيلد (**Cobuild**)، وذخيرة لونجمان - لانكاستر (**Longman/Lancaster**) تستخدم تقنيات المسح الضوئي لتحويل النصوص الورقية إلى مكافئها الرقمي، الأمر الذي سَهّلَ عملية إدخال النصوص إلى ذاكرة الحاسب بأعداد كبيرة. ولهذا، فقد صاحب ذلك ازدياد في حجم تلك الذخائر اللغوية، وقد اقترح سِنكلير عام ١٩٩١ قائلاً:[20] إنه بالإمكان بناء ذخيرة لغوية ذات فائدة وصغيرة الحجم تتراوح بين عشرة أو عشرين مليون كلمة. ولكننا إذا أردنا أن نُقَدِّم وصفًا للنظام اللغوي بأكمله يُعتَمَد عليه، فإن هذا الحجم يظل صغيرًا إلى حد ما، لدرجة أن هناك من يعتقد أن الحدود المُقَيِّدة لكبر حجم النصوص يكاد يكون النقيصة الملازمة لبنية الذخائر اللغوية. على سبيل المثال نجد أن سِنكلير قد سبق له أن أشار في عام ١٩٩١ قائلاً: إننا حتى لو نجحنا في بناء ذخيرة لغوية من مليار كلمة فسوف تظل أمامنا معلومات لا تظهر في قوائم المفردات عن الحالات النادرة لأنواع الكلمات. وفي الحقيقة أن هذا الرأي هو ما يشير إليه قانون زيبف (**Zipf's law**) القائل: إننا لو رمزنا بالرمز"**f**" إلى عدد مرات تكرار الكلمات من الأكبر فالأقل في الجدول التكراري للكلمات، وأخذنا ترتيب كل كلمة في ذلك الجدول من الأقل فالأعلى ورمزنا له بالرمز"**r**" (**Rank**) وكانت:

$$r = 1,2,3,\ldots$$

فسوف نجد أن هناك تناسبًا عكسيًا بين"f"، و"r"، بمعنى أن:

$k = r * f$

أو

$f = k / r$

حيث تشير "k" إلى قيمة ثابتة.

ولا يصعب علينا ملاحظة أن هناك دالة يمكن تمثيلها بخطين بيانيين أحدهما يمثل عدد مرات تكرار الكلمات "f"، والآخر يمثل ترتيب الكلمات "r" في الجدول التكراري. وقد أثبتت الدراسات التي تمت على العديد من الذخائر اللغوية صحة قانون زييف: فهناك عدد قليل من الكلمات التي تتمتع بنسبة تكرار عالية قد غطَّت إجمالي عدد مرات التكرار لمعظم الكلمات المكونة للذخيرة اللغوية، ولوحظ أيضًا أن أكثر من نصف عدد الكلمات في تلك الذخيرة اللغوية لم يظهر إلا مرة واحدة. وأثبتت دراسة أخرى أُجْرِيَت في السنوات الأخيرة أن قانون زييف لا يصلح فقط للاستخدام في دراسة توزيع الكلمات داخل اللغة، وإنما يُستَخدم أيضًا في دراسة توزيع القواعد النحوية في الاستخدام اللغوي. فمن ناحية، وجدنا أن عددًا قليلاً من القواعد النحوية قد غطى إجمالي عدد مرات تكرار معظم الظواهر النحوية الأخرى، ومن ناحية أخرى وجدنا أن هناك العديد من القواعد لم يظهر إلا مرة واحدة داخل الذخيرة اللغوية. والطريف أننا اكتشفنا أن القواعد النحوية كانت تتزايد بزيادة حجم الذخيرة اللغوية. وقد مَثَّلت تلك النتيجة الإحصائية تحديًا كبيرًا أمام إحدى الفرضيات الشهيرة التي طرحها تشومسكي التي تقول: إن قواعد اللغة محدودة في حين أن عدد الجمل داخل اللغة لا نهائي.

ما عدد العلامات التي نحتاجها إذا أردنا أن نقدم وصفًا كاملاً لإحدى اللغات؟ تشير الإحصاءات إلى أن نسبة ٤٠ – ٥٠% تقريبًا من إجمالي عدد الكلمات المكوّنة لذخيرة لغوية يصل حجمها إلى مليون كلمة لا يتكرر غير مرة واحدة. وقد انتبه كلا من رَنْدِل وستوك (Rundle and Stock) [21] عام ١٩٩٢ في حديثهما عن علم المعاجم القائم على الذخائر اللغوية أن كلمة "Break" على الرغم من أنها ظهرت ٨٢٦٧ مرة في ذخيرة لونجمان/لانكستر (Longman/Lancaster) فإن عدد مرات تكرار تلك الكلمة في تركيب مثل: "news breaking" بمعنى "انتشار الأخبار" كان قليلاً إلى الحد الذي لا يمكن معه إعطاء معلومة لمؤلفي المعاجم أن هاتين الكلمتين تؤلِّفان معًا مصطلحًا ينبغي أن يَرِدْ في المعجم. ولكن إذا اعتبرنا أن ظهور الكلمة أو التركيب مرة واحدة غير كافٍ، إذن فكم عدد مرات التكرار التي يمكن أن تُحَدِّد إمكانية النظر بعين الاعتبار إلى مصطلح أو معنى كلمة إذا تكررت داخل الذخيرة اللغوية؟ على سبيل المثال أُجْرِيَت دراسة على ذخيرتي لوب (Lob) وبراون (Brown) فلوحِظ أن ٩٠% من مرات تكرار كلمة "Circumstance" جاء بصيغة الجمع "Circumstances" ومن ناحية أخرى لوحظ أن كلمة من الكلمات التي تتمتع بمعدل تكرار عالٍ في الإنجليزية ككلمة "at" قد ظهرت ٥٥٠٠ مرة مقترنة بكلمات أخرى، وهذا الرقم كبير جدًّا. وبالنسبة إلى علم المعاجم أو أبحاث المفردات والقواعد فإنه كلما تضخمت المعلومات المستخرجة آليًّا من الذخيرة اللغوية زادت الحاجة إلى معالجتها يدويًّا، وهذا ما يصعب التعامل معه، فالكلمة إذا تجاوز عدد مرات اقترانها مع كلمات أخرى ١٠٠٠ مرة، سيمثل ذلك أكبر تحدٍّ يواجه

صناع المعاجم فيما يتعلق بإشكالية تحليل البيانات. ومن هنا، نجد أن علماء اللغة ومؤلفي المعاجم يطالبون بضرورة وجود نوع من البرمجيات والتقنيات الفعالة تساعدهم في أداء أعمالهم من أجل استخلاص أفضل النماذج اللغوية من بين كم كبير من المادة اللغوية. إن هذا ما يحدث عند تحليل المفردات ذات معدل التكرار الأعلى داخل الذخيرة اللغوية. وبناءً على ذلك، يمكننا القول: إن كِبَر حجم الذخيرة اللغوية يتسبب في ظهور بعض الآثار الجانبية عند التحليل اللغوي. وعادةً ما نجد أن تلك الإشكاليات يتم التغلب عليها مباشرة على الحاسب الآلي في أثناء انتقاء العينات. ولكن بالنسبة إلى التراكيب التي تظهر مرة واحدة داخل الذخيرة اللغوية يكون هناك حاجة لأن تُصنَّف داخل قائمة فرعية، ثم نقوم بالبحث عن الأمثلة التي ترد فيها تلك التراكيب في ذخائر لغوية أكبر حجمًا.

وقد أظهرت الدراسات أن ذخيرة يصل حجمها إلى مليون كلمة مثل ذخيرة لوب (**Lob**) أو ويلنجتون (**Wellington**) تحتويان تقريبًا نحو مائة كلمة ورد ذكرها أكثر من ألف مرة. بالإضافة إلى أن هذه القائمة نفسها من الكلمات عندما عُرِضَت على ذخيرة بي إن سي (**BNC**) البالغ حجمها مائة مليون كلمة، وُجِد أن معدل التكرار قد تضاعف بصورة كبيرة ليصل إلى ٨٠٠٠ مرة، وأن هذا العدد (الـ ١٠٠ كلمة) قد غطى ما يقرب من ٩٥% من إجمالي عدد مرات تكرار الكلمات. أما الـ ٥% الباقية من عدد مرات التكرار (التي يصل عددها خمسة ملايين مرة) فقد اقتصرت على حوالي خمسمائة ألف كلمة فقط.

ولذلك إذا أردنا أن ندرس الظواهر الصرفية داخل لغـة مـا، وعلـى الأخص إذا كان هدفنا وضع توصيف دقيق للكلمـات ذات معـدل التكـرار المنخفض وأساليب اقترانها مع الكلمات الأخرى ستكون حاجتنا أكبـر إلـى الذخائر اللغوية كبيرة الحجم. ولكن إذا تعذر التعامل مع البيانات الكبيرة فلن يكون من المفيد في ذلك الوقت الاستعانة بمثل هذا النوع من الذخائر. حتـى إذا استطعنا بناء ذخيرة لغوية كبيرة وجمعنا في داخلها مثل حجم المحيط من المواد اللغوية، فإذا لم يكن تصميم هذه الذخيرة اللغوية متميزًا، لن نتمكن من تقديم وصف لغوي من خلالها. بناءً على هذا المفهـوم يمكننـا القـول: إن الذخيرة اللغوية كبيرة الحجم ليس بالضرورة أن تكون أكثر -تمثيلاً للُغَةٍ ما- دونًا عن الذخيرة الأقل حجمًا. وفي المرحلة الراهنة لا يمكننا أن نقطع بمدى ما يجب أن يكون عليه حجم الذخيرة اللغوية العامة أو المتخصصة. ولـذلك يرى الكاتب ومعه محللو الذخائر اللغوية أنه من الأفضل الاهتمـام بجـودة المادة اللغوية بدلاً من الإفراط في الاهتمام بحجم الذخيرة اللغوية ونطاقها.

ويشير المرجع رقم [15] في قائمة المراجع إلى أننا لو أردنـا دراسـة عَروض اللغة الإنجليزية على سبيل المثال، فإن ذخيرة لغوية تصل إلى مائة ألف كلمة تكفي لهذا الغرض. وإذا أردنا أن نُجْرِي تحليلاً يمكن الاعتماد عليه عن استخدامات الفعل في اللغة الإنجليزية فسوف يمكننا تنفيذ هـذه المَهَمَّـة اعتمادًا على ذخيرة لغوية يصل حجمها إلى خمسمائة ألف كلمة؛ وأن دراسة التراكيب النحوية والكلمات ذات معدل التكرار الأعلى تحتاج غالبًا إلى ذخيرة تتراوح ما بين خمسمائة ألف ومليون كلمة. كما يمكننا أن نُجْرِي مقارنة بين

نتائج تحليل المادة اللغوية في ذخيرة كبيرة الحجم مثل ذخيرة بـي إن سـي (BNC) وبين نتائج التحليل في ذخائر صغيرة مثل لوب (LOB)، وآي سي إي (ICE) لكي نتوصل إلى حل الإشكاليات المتعلقة بحجم الذخيرة اللغويـة ومدى مرجعيتها. وبهذه الطريقة يمكننا التوصل إلى معرفة حجـم الـذخيرة اللغوية الذي يناسب كل غرض من الأغراض البحثية. وبصورة عامة يمكننا القول: إنه مهما كان حجم المادة اللغوية فغالبًا ما تزداد نسبة الاعتماد علـى نتائج التحليل اللغوية لتلك المادة والاعتراف بها كلما ازداد عـدد العينـات اللغوية التي تغطي قطاعًا عريضًا من التخصصات داخل اللغة.

٢- عدد العينات

القضية الأخرى المتعلقة بحجم الذخيرة اللغوية هـي عـدد العينـات المأخوذة داخل الذخيرة. ففي الجيل الأول من الذخائر اللغوية مثـل ذخيـرة لوب (LOB) وذخيرة براون (BROWN) وصلت العينات المُمَثِّلـة داخـل هاتين الذخيرتين إلى ٥٠٠ عينة، وكان حجم النصوص في كل عينة لا يقل عن ٢٠٠٠ كلمة. من بينها عدد من العينات كانت تُمَثِّل نصًا كـاملاً، أمـا غالبية العينات فكانت عبارة عن جزء من نص. ونظرًا إلى أن النـصوص التي تُقْتَطَف من مقدمة الكتب وخاتمتها تتميز أنها تعبر تعبيرًا دقيقًـا عـن الخصائص اللغوية المختلفة، لذلك فإن الذخائر اللغوية التي تُبْنـى بأسـلوب العينات النصية لا يمكن أن تُعَبِّر عن الإطار العام للغة إلا إذا كانـت تلـك العينات تشمل جميع أنواع النصوص اللغوية، وإذا لم يراعَ ذلك في اختيـار

العينات فسوف تَضُرُّ الذخيرة بالملامح العامة للغة. وعلى العكس من ذلك إذا تم بناء الذخيرة على أساس النصوص الكاملة فسوف يؤدي ذلك إلى نقص عدد العينات وقلة أنواع النصوص المُمَثَّلَة داخل الذخيرة. ومن ثم يؤثر ذلك في قدرة الذخيرة على أن تكون مُمَثِّلَة للغة. ولا يمكن التغلب على تلك المشكلة إلا بكبر حجم الذخيرة اللغوية.

وقد قام بيبر **(Biber:1993)**[22] بعمل دراسة عن توزيع عشر خصائص لغوية داخل ذخيرتي لوب **(LOB)** وإل إل سي **(LLC)** وذلك من خلال ٥٥ زوج من العينات اللغوية، وقد اختار هذه العينات من مساحة عريضة من النصوص الشفهية والتحريرية. حيث كان يختار كل زوج من العينات في إطار ألفي كلمة من النصوص مُوَزَّعَة بالتساوي على الذخيرتين وفي تخصص واحد، ثم يدرس الاختلافات اللغوية بين الاثنين من خلال تلك العينات. وكانت النتيجة التي توصل إليها هي أنه يكفي وجود ما بين ٢٠٠٠ - ٥٠٠٠ كلمة حتى يمكن أن تكون العينة مُعَبِّرَة عن الخصائص اللغوية لنصٍّ ما. بالإضافة إلى أنه يَعْتَقِد أن عدد ٢٠ - ٨٠ عينة لغوية مختلفة تكفي لدراسة الاختلافات اللغوية الشائعة.

الفصل الثاني

بناء ذخيرة لغوية

يشير المرجع رقم [23] في قائمة المراجع إلى أن وضع تعريف للذخيرة اللغوية، وتحديد إجمالي حجم النصوص المختارة وأنواعها، ونسبة تمثيل كل نوع من أنواع النصوص داخل الذخيرة من أصعب الأعمال التي تواجه عالم اللغة في أثناء بناء الذخيرة اللغوية. ويرى بعضهم أن هذا العمل ينبغي أن يُوكَل إلى العلماء المتخصصين في علم اللغة الاجتماعي للقيام به. أما علماء اللغة فينبغي أن تقتصر بحوثهم على إجراء تحليل للأمثلة اللغوية التي تحتويها الذخيرة اللغوية وتوصيفها. ولكن واقع الأمر حاليًا يُشير إلى أن علماء اللغة وخبراء الحاسب هم الذين يقومون بأعمال التصميم الخاصة بتوزيع العينات داخل الذخيرة اللغوية، ومما لا شك فيه أن هذا العمل لا يخلو من صعوبة بالغة.

إن أول شيء يتم التفكير فيه عند بناء ذخيرة لغوية هو الغرض الذي تُبْنى من أجله تلك الذخيرة، وهل هي ذخيرة للأغراض العامة أم الخاصة. تُقَدِّم ذخائر الأغراض العامة كمية كبيرة من الأمثلة اللغوية للبحث اللغوي في جميع المجالات، أما ذخائر الأغراض الخاصة فتُسْتَخدم في إنتاج التطبيقات القائمة على معالجة اللغات الطبيعية، وتتصف بأنها تُمَثِّل مجالا لغويًا محددًا؛ ولذلك ينظر البعض إلى هذا النوع من الذخائر اللغوية على أنه إحدى مراحل

ابتكار تلك التطبيقات وتطويرها. ويتمحور الغرض من هذا الباب حول التعريف بالذخائر اللغوية التي تُصَمَّم لأغراض الاستخدام العام.

أولاً: مصادر المادة اللغوية

تُسْتَخْدَم في ذلك وسائل إدخال البيانات المتاحة بالفعل، التي تتعدد أشكالها في جمع المواد اللغوية إما عن طريق تحويل النصوص الورقية إلى نسخة رقمية أو عن طريق الاستعانة بالنصوص الرقمية الموجوده بالفعل.

١- تحويل النصوص الورقية إلى نُسَخ رقمية

هناك طريقتان للمساعدة في تحويل النصوص الورقية إلى نسخ رقمية[١]، أولاهما:

أ) طريقة المسح الضوئي

زادت في الآونة الأخيرة تقنيات التمييز الضوئي للحروف المطبعية بصورة تدريجية؛ ولذلك فإن كميات كبيرة من النصوص المكتوبة بالحروف المطبعية يمكن أن تعتمد على تلك التقنيات حتى يمكن تحويلها إلى نسخ رقمية. وتجنبنا هذه الطريقة اللجوء إلى لوحة المفاتيح لإعادة كتابة محتويات

---

(١) الجدير بالذكر أن علماء اللغويات الحاسوبية العرب لم ينجحوا حتى الآن في حل تلك المشكلة بالنسبة إلى اللغة العربية، وجميع المحاولات المطروحة في هذا المجال لا ترقى لمستوى الاستخدام؛ بسبب القصور الملحوظ في دراسات المعالجة الآلية لمنظومة الكتابة باللغة العربية. وعلى الجانب الآخر نجد أن مشكلات تحويل الكتابة الصينية بشقيها المطبوع واليدوي إلى مكافئ إلكتروني قد تم حلها بشكل كامل من قبل العلماء الصينيين بأنفسهم مع بدايات القرن الحادي والعشرين، على الرغم من الصعوبات الهائلة التي اعترضتهم بسبب طبيعة اللغة الصينية التي تعتمد على الرموز، الأمر الذي أهل اللغة الصينية للدخول إلى مستويات عصر المعلوماتية بكفاءة غير مسبوقة. (المترجم)

الكتب المراد تسجيل محتوياتها في الذخيرة اللغوية. على سبيل المثال يُباع في الأسواق الصينية حاليًا برمجية تتّبع تلك التقنية في إدخال النصوص المكتوبة بالحروف المطبعية إلى الحاسب وتحويلها إلى نسخة رقمية. وقد تم تطوير هذه البرمجية بمعرفة قسم علوم الحاسب وتقنياته بجامعة تشين خوا الصينية[24]. ولكن ينبغي الانتباه إلى أن استخدام هذا الأسلوب في تمييز النصوص المطبوعة لا يخلو من بعض الأخطاء، ويحتاج الأمر إلى إجراء تصويبات يدوية لتصحيح تلك الأخطاء التي تقع من البرمجية في أثناء عملية تمييز النصوص وإدخالها إلى وسائط التخزين الإلكترونية داخل الحاسب.

ب) إدخال النصوص إلى الحاسب يدويًا عن طريق لوحة المفاتيح

يتم اللجوء إلى استخدام هذا الأسلوب مع أنواع النصوص التي لا يمكن استخدام أسلوب التمييز الضوئي لها. وهذا النوع من النصوص يشمل نصوص الفاكسات، والمذكرات الشخصية، والتسجيلات الصوتية وغير ذلك. ففي بعض الأحيان يكون استخدام أسلوب التمييز الضوئي للنصوص غير فعال بالدرجة الكافية؛ حيث يحتاج إلى وقت كبير لإجراء تصويبات يدوية على النسخة التي تم تمييزها، لدرجة أن يكون من الأفضل الاعتماد على شخص ذي كفاءة عالية في كتابة النصوص يدويًا على الحاسب لتحويل هذا النوع من النصوص إلى نسخ رقمية. في البدايات الأولى لإنشاء الذخائر اللغوية باللغة الصينية، كان المتخصصون يلجئون إلى مثل هذا الأسلوب في العمل؛ ففي بداية ثمانينيات القرن العشرين، قام قسم علوم الحاسب بجامعة شان شي بتكليف من اللجنة القومية للغة الصينية بعمل دراسة إحصائية عن

معدل تكرار الرموز داخل اللغة الصينية عن طريق بنـــاء ذخيـــرة لغويـــة باستخدام هذا الأسلوب[25]؛ حيث تم تكليف بعض الأشخاص المهرة بكتابـــة محتويات جريدة الشعب اليومية خلال شهر وتحويلها إلى نسخة إلكترونيـــة، وبعد ذلك تم استخدام الحاسب في إنهاء الأعمال الخاصة بإحـــصاء معـــدل تكرار الرموز داخل تلك الذخيرة.

٢ــ استخدام النصوص الرقمية الموجودة بالفعل

أصبح من الشائع الآن توافر نسخ إلكترونية من الجرائـــد والمجـــلات والكتب وغير ذلك من المواد المطبوعة، ويُعْتَبَر اللجـــوء إلـــى مثـــل هـــذه النصوص الوسيلة المباشرة لبناء الذخائر اللغوية، وفي تلك الحالة ينحـــصر عمل مصممي الذخائر على القيام بالتهيئة اللازمة لهذه النـــصوص؛ حيـــث يقومون بتحويلها إلى النمط الرقمي المناسب لبنية النصوص داخل الـــذخيرة اللغوية بما يُمَكِّن من التعامل معها فيما بعد.

وفي الحقيقة أن عددًا كبيرًا من مشروعات بناء الذخائر اللغوية يتم من خلال الجمع بين الأساليب السابقة في إدخال النصوص إلى ذاكرة الحاســـب وتحويلها إلى نسخة رقمية؛ لأن كلا منها يتناسب مع طبيعة كل نـــوع مـــن أنواع النصوص المتاحة، التي تُعْتَبَر المادة الخام لبناء الذخيرة اللغوية. فعلى سبيل المثال، نجد أن النصوص المكتوبة بخط اليـــد والنـــصوص المُـــسَجَّلَة صوتيًا يناسبها استخدام لوحة المفاتيح لتحويلها إلى نصوص إلكترونية. وعلى العكس من ذلك، نجد أن العديد من نسخ الجرائد والمجلات أصبحت تتـــوافر في شكل إلكتروني، أما الكتب المطبوعة بالأسلوب التقليدي فيُفَضَّل استخدام

برنامج مسح ضوئي ذي كفاءة عالية كوسيلة سريعة وفعالة لتحويل نصوص تلك الكتب إلى المكافئ الرقمي لها.

ثانيًا: تصميم الذخيرة اللغوية

مرحلة تصميم الذخيرة اللغوية تشمل تحديد حجمها، ونطاق النصوص المُخزَّنة بداخلها، والنسبة التي يشغلها كل نوع من النصوص من الحجم الكلي للذخيرة، وبنية التصنيفات التي تندرج تحتها تلك النصوص، التي تُمكِّن من البحث داخل الذخيرة، وغير ذلك من الموضوعات التي تَمَس عملية جمع النصوص. ويقدم المرجع رقم[23] من قائمة المراجع وصفًا دقيقًا لما يتعلق بتصميم الذخائر اللغوية، وسوف نقدم فيما يلي شرحًا مبسطًا لبعض الموضوعات المحورية التي تستحق التركيز مضافًا إليها بعض آراء المؤلف.

١- توزيع المادة اللغوية داخل الذخيرة

من ناحية الجوهر، تُعتَبَر الذخيرة اللغوية كيانًا يضم في داخله جميع أنواع النصوص. ولذلك فإن مسألة اختيار أنواع النصوص التي يتم التعامل معها من العناصر التي لها أكبر الأثر في كفاءة الذخيرة اللغوية. وسوف نتحدث فيما يلي عن بعض النقاط التي ينبغي أن يكون لها الأولوية في التفكير عند بناء الذخائر اللغوية:

أ) المستويان الشفهي والتحريري للغة

عند بناء الذخيرة اللغوية ينبغي أن نُحَدِّد سابقًا ما إذا كانت ذخيرة للنصوص التحريرية أم للنصوص الشفهية، أم أنها تجمع بين النوعين.

والجدير بالذكر أن هناك العديد من الذخائر اللغوية لا تضم في داخلها إلا نماذج لنصوص تحريرية. وهذا من شأنه أن يتسبب في كون الذخيرة غير معبرة عن اللغة ولا تعكس الواقع العملي للاستخدام اللغوي؛ حيث يعتقد العديد من علماء اللغة أن اللغة الشفهية أكثر قدرة على التوجيه والإرشاد في عمل الدراسات الأساسية عن التكوين اللغوي، وأنه لا يوجد أي شكل من النصوص التحريرية يمكن أن يلعب هذا الدور بصورة كاملة.

إلا أن عمليات جمع النصوص الشفهية أصعب بكثير من جمع النصوص التحريرية. وعلى الرغم من عدم صعوبة جمع بعض أشكال اللغة الشفهية مثل سيناريوهات السينما ونصوص المسرح وسجلات المؤتمرات ومرافعات القضايا داخل المحاكم والنشرات التليفزيونية، فإن اللغة التي تمثلها هذه الأنماط اللغوية قد تَمَّت عليها بعض التعديلات، ومن المؤكد أننا سوف نعثر في متنها على بعض آثار المعالجة الاصطناعية؛ ولذلك لا يمكن لهذه النصوص أن تُعَبِّر عن الروح اللغوية التي يتسم بها الحوار الطبيعي. وغالبًا ما نُطلِق على هذا النوع من اللغة الشفهية مصطلح اللغة الشفهية القياسية.

عند بناء الذخائر اللغوية الصينية تكون النسبة الأكبر هي النصوص التحريرية؛ مثل الجرائد والكتب وغير ذلك. وتقل نسبة المكوِّن الشفهي داخل الذخائر الصينية وبخاصة ذلك الجزء المعبر عن الفوارق الطبيعية بين الأشخاص.

ب) المستوى الرسمي والمستوى الأدبي من اللغة

المادة اللغوية يمكن اختيارها من بين العديد من الأشكال اللغوية، فقد تكون تلك المادة في شكل رسمي أو غير رسمي، أو قد تكون لغة أدبية أو

لغة عادية. فاللغة الرسمية يَسْهُل الحصول عليها مقارنة باللغة غير الرسمية. أما اللغة الأدبية فيَسْهُل رصدها مقارنة باللغة العادية، أما الرسائل الكتابية واليوميات القصيرة غير الرسمية التي يَغْفُلها الناس فتُعْتَبَر الوجه الحقيقي الذي يُمَثِّل الاستخدام اللغوي في إطاره الأشمل. وبالمثل فلا توجد حاجة لأن تحتوي الذخيرة اللغوية جميع الأعمال الأدبية المعاصرة؛ حيث إن نسبة قليلة منها تفي بالغرض.

ج) مدى مرجعية اللغة

الوظيفة الأساسية للذخيرة اللغوية هي إعطاء الإمكانية للإبحار داخل اللغة لاكتشاف جوهرها والعناصر القياسية المعبرة عنها. فلو كانت غالبية النصوص في الذخيرة تم اختيارها من أعمال أديب واحد، فمن المؤكد أن تلك الذخيرة سوف تتمحور بصورة كبيرة حول التعبير عن الخصائص الإبداعية لأسلوب هذا الكاتب، وبذلك تفقد قيمتها في إجراء البحوث اللغوية المعبرة عن الخصائص العامة للغة.

وعلى المنوال نفسه نجد أن كل كاتب من كُتَّاب الجرائد والمجلات له أسلوب مختلف في الكتابة، ومن هنا فإن جمع أكبر قدر من النصوص على مستوى جميع التخصصات من الأمور المفيدة جدًّا للبحث اللغوي. وإذا أردنا أن نجعل الذخيرة اللغوية أكثر قربًا من واقع اللغة فينبغي لنا أن نضم في داخلها أكبر قدر من أعمال الكُتَّاب في مختلف التخصصات والمجالات.

د) الفترة الزمنية للنصوص

غالبية الذخائر اللغوية تسعى إلى تغطية فترة زمنية محددة، وبما يجعل المادة اللغوية المُسَجَّلَة داخل الذخيرة اللغوية تُعَبِّر عن الواقع اللغوي للغة في

تلك الفترة الزمنية. وهذا النوع من الذخائر اللغوية يطلق عليه اسم الذخائر اللغوية المُمَثِّلة لزمن محدد. وعلى العكس من ذلك، هناك نوع آخر من الذخائر اللغوية التاريخية تنشأ بغرض عمل سلسلة من الأبحاث عن التغيرات التي تطرأ على اللغة من ناحية المفردات والبناء الهيكلي لها بصورة تاريخية؛ وذلك عن طريق جمع عدد من الأعمال في فترات تاريخية مختلفة.

٢ــ نطاق المادة اللغوية

عند تحديد نطاق المادة اللغوية يكون هناك حاجة إلى التفكير في العناصر التالية:

أ) حجم الذخيرة اللغوية

هذه النقطة من أهم القضايا التي يهتم بها مصممو الذخائر اللغوية في بداية العمل في إنشاء أي ذخيرة. ففي الفترة الأولى من تاريخ البحوث المتعلقة بالذخائر اللغوية لم تكن هناك وسيلة لبناء الذخائر اللغوية إلا عن طريق التسجيل اليدوي للمادة اللغوية؛ حيث كان من الصعب وجود مكافئ رقمي للنصوص اللغوية. ولذلك كان حجم الذخائر آنذاك صغيرًا إلى حدٍّ بعيد. وتحت تأثير تلك الظروف، كان من الواجب توخي الدقة عند تصميم مخطط توزيع المادة اللغوية حتى تكون الذخيرة اللغوية معبرة عن اللغة بصورة شاملة. وإذا نظرنا إلى الأمر نظرة بعيدة المدى يمكننا التنبؤ بأن الذخائر اللغوية سوف يتأثر حجمها تبعًا لتطور تقنيات الحاسب. وهذا ما يخبرنا به قانون زييف من أن نسبة توزيع الكلمات المتكررة داخل الذخيرة اللغوية غالبًا ما تتناقص بصورة كبيرة تبعًا لتناقص عدد مرات تكرار تلك

الكلمات داخل النصوص، على سبيل المثال، تلك الدراسة التي تمت على اللغة الإنجليزية التي تشير إلى أن كلمة "the" التي سَجَّلت أعلى عدد لمرات التكرار داخل اللغة الإنجليزية، جاء عدد مرات تكرارها ضعف عدد مرات تكرار كلمتي "of"، و"and" اللتين تأتيان في المرتبة الثانية من حيث التكرار، وبعد ذلك يتناقص عدد مرات تكرار الكلمات بصورة سريعة إلى أن نصل إلى كلمة "be"، التي تحتل المركز التاسع عشر من حيث الكلمات الأكثر تكرارًا في اللغة الإنجليزية، فنجد أن نسبة تكرار هذه الكلمة يمثل ١٠% فقط من نسبة تكرار كلمة "the"، أما الكلمة التي تحتل المركز ٨٤ وهي كلمة "two" فتشغل ٥% فقط من نسبة تكرار كلمة "the".

ولذلك إذا أردنا أن نستعين بمنهجية الذخائر اللغوية في عمل استقراء لمفردات لغة ما، وأردنا أن تغطي دراستنا أكبر قدر من مفردات تلك اللغة، فعلينا توسيع من حجم الذخيرة قدر استطاعتنا.

ب) حجم العينات اللغوية

وأخيرًا، ما الحجم الذي ينبغي أن تكون عليه كل عينة لغوية؟ هذا هو السؤال الذي يتكرر دائمًا عند مصممي الذخائر اللغوية، ولكل منهم رأيه الخاص، على سبيل المثال، نجد أن ذخيرة لوب LOB لا تقل كل عينة فيها عن ٢٠٠٠ كلمة، وعلى الرغم من أن هذه الطريقة قلَّدَها العديد من العاملين في بناء الذخائر اللغوية، فإن هناك عددًا من العلماء وجهوا نقدًا لهذه الطريقة؛ حيث يعتقدون أن النصوص بهذا الحجم لا تكفي للتعبير عن الخصائص اللغوية للنص الأصلي. على سبيل المثال رسائل الأخبار القصيرة

(مثل الجمل الإخبارية) نجد أن أسلوبها يختلف عن أسلوب كتابة التقارير الإخبارية؛ ولكن نظرًا إلى أن حجم النصوص التي تمثل رسائل قصيرة لا يكفي لأن يكون في حدود ٢٠٠٠ كلمة، فلا يمكن أن تضم الذخيرة اللغوية هذا النوع من النصوص؛ ومن ثَمَّ لا يُتاح للذخيرة اللغوية أن تغطي مثل هذا النوع من النصوص الذي يقدم لنا الخصائص اللغوية للرسائل الإخبارية القصيرة.

وعلى الجانب الآخر نجد أن اختيار عينة في حدود ٢٠٠٠ كلمة للتعبير عن الخصائص اللغوية لرواية طويلة لا يمكن أن تفي بهذا الغرض؛ ولذلك فمن أجل أن نعوض ذلك النقص ونتلافى هذا العيب علينا اختيار النص بأكمله لكي يدخل في متن الذخيرة اللغوية حتى يمكن تجنب انحصار العينات في بعض فقرات النص الأصلي.

ولذلك، فعند إتاحة الظروف يكون من الأفضل اختيار النصوص بأكملها. وهذه الطريقة لا يُخشى معها الوقوع في مشكلة التفاوت بين أساليب الأجزاء المختلفة للنص.

إن اختيار النصوص بأكملها أفضل من الاكتفاء باختيار عينات منها؛ لأن النص الكامل من شأنه أن يُقَدِّم معلومات شاملة تساعد على البحث اللغوي، ومع هذه الطريقة لا يكون داعٍ هناك للقلق من عدم الالتزام بالخُطَّة المُتَّبَعَة في بناء الذخيرة اللغوية. وأهم ما في الأمر هو ضمان تصميم برنامج

جيد لإدارة البيانات داخل الذخيرة اللغوية بما يتيح الفرصة أمام علماء اللغة للحصول على ما يبغون التوصل إليه عند الحاجة.

٣- تصميم برنامج التخزين وسجلات الحفظ

إن الهدف من إنشاء ذخيرة لغوية هو استغلال الحقائق اللغوية واستخدامها في إجراء البحث اللغوي. ولذلك يُعْتَبَر الالتزام بمعايير الدقة في تصميم وسيلة تخزين المادة اللغوية وإشارات ترميزها من الأعمال المهمة جدًّا خلال عملية إنشاء الذخائر اللغوية. فلو أخذنا اللغة الإنجليزية مثالاً، نجد أن تخزين ذخيرة لغوية سعتها مليون كلمة تحتاج إلى مساحة تخزين تتراوح بين ٨ و١٠ ميجا بايت. وإذا أضفنا علامات الترميز النحوية لمحتويات هذه الذخيرة سنحتاج من ٣ إلى ٥ ميجا بايت إضافية، أما إذا أضفنا رموز التحليل النحوي فسوف نحتاج إلى ٣٠ ميجا بايت أخرى. إن عملية تخزين المادة اللغوية في متن الذخائر اللغوية لم تَعُد مشكلة؛ نظرًا إلى تطور تقنيات الحاسب الآلي، فعلى سبيل المثال يمكننا تخزين معلومات تصل إلى مائة ميجا بايت على أسطوانة ليزر واحدة. إن الذخيرة اللغوية لا تكتسب قيمتها إلا من المادة اللغوية المُخَزَّنَة والمعلومات المتعلقة بها. ولذلك فعند بناء ذخيرة لغوية علينا أن نبدأ بتصميم أسلوب للاستعلام داخل الذخيرة بشكل منظومي، يضمن لنا ارتباط النصوص التحريرية داخل الذخيرة بمكافئها الأصلي سواء كان صوتًا أو نصًّا. وبالإضافة إلى ذلك، ينبغي حفظ فهارس الملفات وجميع الملفات في نسخ إضافية غير تلك التي تُجْرَى عليها عمليات التحليل والدراسة.

ومن ناحية أخرى نجد أننا نلجأ إلى مصادر مختلفة للحصول على المادة اللغوية، فلو اتبعنا أساليب مختلفة لتكويد تلك المواد اللغوية وترميزها،

فسوف يؤدي ذلك إلى حدوث خلل كبير في إدارة الحاسب وسيطرته على تلك المواد اللغوية. ولكي نصل إلى أسلوب موحَّد في إدارة تلك المواد اللغوية نحتاج إلى استغلال ملامح النصوص واستخدامها كعلامات للترميز، وأكثر تلك الملامح التي نلجأ إليها هي علامات الترقيم وأرقام السطور داخل النص والعلامات الخاصة ببدايات الأبواب والفصول والفقرات في النص، وغير ذلك. وإذا لم تكن تلك العلامات مُوَحَّدَة داخل الذخيرة اللغوية فلن يتمكن الحاسب من التوصل إلى معلومات ذات فائدة فيما يتعلق ببنية النصوص داخل الذخيرة؛ ومن ثَمَّ لا يمكنه تمييز النصوص وأكوادها المختلفة، الأمر الذي ينشأ عنه مخرجات وبيانات خاطئة.

في ثمانينيات القرن العشرين توصل العاملون في مجال النشر إلى معيار قياسي لتكويد النصوص إلكترونيًّا من أجل تجنب وجود تكرار غير مرغوب فيه في أثناء عمليات إدخال النصوص، ونتيجة لوجود لغة العلامات(١) التي أُطلِقَ عليها اختصارًا اسم SGML (The Standard Generalized Markup Language)، أصبح المتخصصون في هذا المجال يستخدمونها في عمل تكويد النصوص الإلكترونية.

---

(١) يُستخدم مفهوم لغة العلامات (Markup Language) في مجال تكنولوجيا المعلومات للإشارة إلى ثنائيات من الكلمات والرموز توضع بينها المكونات النصية لصفحة الويب، بالإضافة إلى مجموعة من العلامات تؤدي إلى عرض مكونات الصفحة وفقًا لمواصفات معينة، مثل نوع الخط وحجمه ولونه، وهل تُعرَض البيانات في شكل نصي أم في شكل جدولي، وشكل محاذاة الأسطر، وما إلى ذلك من مواصفات تحرير النصوص. وتنتمي لغة الـ SGML ، والـ HTML، والـ XML إلى العائلة اللغوية نفسها، وتستخدم بصورة أساسية في تصميم صفحات الويب.(المترجم)

وبقي أن نشير إلى أنه قبل إجراء إعداد أنماط النصوص وتوحيدها بعد تجميعها داخل الذخائر اللغوية علينا أن نحصل على حقوق استغلال تلك النصوص من أصحابها الأصليين حتى لا تواجهنا مشكلة حقوق التأليف والنشر.

٤- حماية الذخيرة اللغوية

بعد أن تُبْنى الذخيرة اللغوية دائمًا ما نكتشف بها العديد من الأخطاء التي تحتاج إلى تصحيح، أو حتى إجراء تحسينات على بنيتها؛ ولذلك تكون هناك حاجة إلى إجراء صيانة يومية للذخيرة اللغوية وتطوير دائم لها. وبهذه الطريقة، يمكننا التوافق مع مستجدات الحاسب الآلي من حيث العتاد والبرمجيات، ومن حيث احتياجات المستخدم أيضًا، هذا بالإضافة إلى أن تزايد الاهتمام بأنظمة فهرسة الذخائر اللغوية وأدوات تحليلها ومعالجتها، قد أدى إلى وجود حاجة متزايدة لحماية الذخائر اللغوية.

الفصل الثالث

أنواع الذخائر اللغوية

إن تنوع أشكال الذخائر اللغوية حقيقة لا يستطيع أحد إنكارها؛ ولكن لَمْ يتوصل أحد حتى الآن إلى تسمية موحدة لكل نوع من هذه الأنواع. وقد سبق أن وضع دونالد ووكر (Donald Walker)[26] المصطلحات الأربعة التالية لتسمية الأنواع المختلفة من الذخائر اللغوية، إلا أن هذه القضية ما زالت مثيرة لجدل لم يُحْسَم حتى الآن.

١ــ الذخيرة المختلطة (Heterogeneous)

هذه أبسط وسيلة لجمع المادة اللغوية داخل ذخيرة لغوية؛ حيث يسعى المصممون قدر الإمكان إلى جمع كل ما يمكن جمعه من أنواع النصوص المختلفة، ولا يكون هناك أية قواعد سابقة فيما يتعلق باختيار المادة اللغوية، وتُعْتَبَر ذخيرة أيه سي إل/ دي سي آي **ACL/DCI** من الذخائر اللغوية التي تندرج تحت هذا النوع. وتتفق مع هذا النوع ذخيرة أو تي أيه **OTA** للمستندات التابعة لجامعة أكسفورد، حيث جَمَعت هذه الذخيرة بين أنواع مختلفة من المستندات دون أي تغيير في الشكل الأصلي للمستند.

٢ــ الذخيرة المتجانسة (Homogeneous)

وهي عكس النوع السابق، وتُعْتَبَر ذخيرة تيبوتر **Tipoter** التابعة للحكومة الأميركية خير مثال على هذا النوع من الذخائر؛ حيث جَمَعَت في

طياتها النصوص المتعلقة بالشئون العسكرية فقط مثل أعطال التجهيزات وتوقعات الأخطار وغير ذلك من الأمور المتعلقة بالشئون العسكرية. هذا وتُعْتَبَر الذخائر اللغوية التي تُصمم من أجل تخزين أعمال كاتب معين من هذا النوع أيضًا.

٣- الذخيرة المنظومية (Systematic)

الغرض من إنشاء هذا النوع من الذخائر اللغوية ضمان جمع المادة اللغوية بما يُغَطِّي إطارًا لغويًّا واسعًا؛ بحيث تمثل الذخيرة اللغوية أكبر مساحة من اللغة. وتُعْتَبَر ذخيرة براون والذخيرة القومية الإنجليزية BNC من الذخائر التي تَحْتَرِم هذه القاعدة في جمع المادة اللغوية. وعند بناء هذا النوع من الذخائر اللغوية يتم الاهتمام بقضايا الثبات والتغير والمرجعية والتوازن، بالإضافة إلى القضايا المتعلقة بنطاق تغطية المادة اللغوية داخل الذخيرة.

٤- ذخيرة الاستخدام المتخصص (Specialized)

يُخَزَّن بتلك الذخيرة كل أنواع الذخائر المتخصصة، مثل ذخيرة العلوم الإنسانية بأميركا الشمالية وذخيرة لغة الأطفال Childes التابعة لجامعة كارنيجي مِيلون Carnegie Mellon University.

الفصل الرابع

تعريف بالذخائر اللغوية خارج الصين

الهدف الأساسي من علم الذخائر اللغوية هو وضع تصور عام لجميع الحقائق والقوانين اللغوية في اللغة الطبيعية؛ وذلك عن طريق عمل دراسات واقعية واسعة النطاق على المواد اللغوية. وقد بدأ العالَم الغربي في إنشاء الذخائر اللغوية مع بداية ستينيات القرن العشرين. وعلى مدى ثلاثين عامًا، تمكنت العديد من الحكومات والمؤسسات الاستثمارية والتجمعات العلمية على مستوى العالم من إنشاء أو السعي إلى إنشاء ذخائر لغوية على جميع الأشكال. وسوف نُعَرِّف القارئ في هذا الفصل بأهم تلك الذخائر الغربية وأكثرها تأثيرًا.

أولاً: ذخيرة إس إي يو SEU

في عام ١٩٥٩، أطلَق عالم اللغة الإنجليزي راندولف كويرك Randolph Quirk مشروعًا أسماه "دراسة في استخدامات اللغة الإنجليزية" (The Survey of English Usage)[27] وقد أُطلِق على هذا المشروع اختصارًا اسم SEU؛ حيث قام كويرك في إطار خُطَّة محددة بجمع كمية كبيرة من المواد اللغوية في تخصصات مختلفة، بالإضافة إلى أنه استغل الحاسب الآلي في تخزين المواد التي جَمَعَها وتصنيفها. وقد كانت هذه هي

المرة الأولى في تاريخ الذخائر اللغوية التي يتم فيها اللجوء إلى حاسب آلـي من النوع العملاق في بناء ذخيرة لغوية. وقد جَمَعَت ذخيرة **SEU** مـا بـين النصوص التحريرية والنـصوص الـشفهية فـي العديـد مـن المجـالات والتخصصات.

والجدول التالي يوضح نسبة تمثيل النصوص بأنواعها المختلفة داخـل تلك الذخيرة.

جدول (٢-١): هيكل المواد اللغوية في ذخيرة SEU

أصول تحريرية (١٠٠ نص)		
مطبوعات (٤٦)	**مواد غير مطبوعة (٣٦)**	**مواد شفهية (١٨)**
علوم إنسانية (٦)	مسودات كتب (٦)	نصوص مسرحية (٤)
علوم طبيعية (٧)	أخبار (٦)	خُطَب رسمية (٣)
تعليم (٦)	مراسلات اجتماعية خاصة (٤)	نشرات إخبارية (٣)
جرائد إخبارية عادية (٤)	مراسلات اجتماعية عادية (٤)	حوارات إخبارية (٤) خواطر (٢)
جرائد إخبارية متخصصة(٤)	مراسلات اجتماعية في شكل مشاحنات (٤)	حكايات (٢)
مؤلفات أدبية (٤)	مراسلات غير رسمية عادية(٤)	×
قانون (٣)	مراسلات غير رسمية في شكل مشاحنات (٤)	×
نقد (٥)	مذكرات (٤)	×
قصص ونثر (٧)	×	×
أصول شفهية (١٠٠ نص)		
خُطَب مُعَدّة سابقًا (٦)	حوارات خاصة غير علنية (٢٤) مشاحنات غير علنية (١٠)	
خُطَب ارتجالية (١٠) تعليقات رياضية (٤)	حوارات خاصة مُعْلَنة (٢٠) مشاحنات مُعْلَنة (٦)	
متفرقات (٤)	حوارات تليفونية خاصة (١٠) مشاحنات تليفونية (١٢)	

ويمكننا ملاحظة أن تلك الذخيرة تحتوي ٢٠٠ نــص تــم توزيعهــا بالتساوي على كل من المستوى التحريري والمستوى الشفهي مــن اللغــة. ويحتوي كل نص حوالي ٥٠٠ كلمة تقريبًا. ويصل إجمالي عــدد الكلمــات داخل الذخيرة كلها إلى حوالي مليون كلمة، ويغطي محتواها جميع طبقــات المجتمع بمستوياته اللغوية المختلفة. وإذا نظرنا نظرة عامة إلى تاريخ تطور الذخائر اللغوية لوجدنا أن ذخيرة إس إي يو **SEU** التي أنشأها كويرك تُعْتَبَر عملا عظيمًا سواء من حيث وجهة النظر البحثية أو من حيث منهجية التنفيذ. وقد اعتُبِرَت هذه الذخيرة اللغوية بمثابة فاتحة طريق جيدة أمام علم الــذخائر اللغوية وقَدَّمَت أفكارًا علمية جديدة في مجال البحث اللغوي.

ثانيًا: ذخيرة براون

في ستينيات القرن العشرين، أنشأ كلٌّ من فرانسيز **Francis** وكوشيرا **Kucera** في جامعة براون الأميركية أول ذخيرة قياسية على مستوى العــالم تقوم بجمع العينات اللغوية طبقًا لقواعد منظومية، وهي ذخيرة براون اللغوية. وقد كان الغرض الأساسي من إنشاء تلك الذخيرة هــو دراســة الإنجليزيــة الأميركية المعاصرة. ويصل حجم تلك الذخيرة إلى مليون كلمة.

وقد تم جمع المادة اللغوية لهذه الذخيرة من النصوص العامة التي كتبها الأميركان في عام ١٩٦١. وتغطي المادة اللغوية خمسة عشر موضوعًا من خلال خمسمائة عينة. وكل عينة لا يقل عدد كلماتها عن ٢٠٠٠ كلمة. وقــد اعتمدت دار النشر التابعة لجامعة براون على الدراسات الإحــصائية التــي

أُجْرِيَت على تلك الذخيرة في إصدار معجم عن معدل تكرار الكلمات في اللغة الإنجليزية[28] وكان ذلك في عام ١٩٦٧. وفي سبعينيات القرن العشرين، صمم كلٌّ من جرين (Greene)، وروبين (Rubin) برنامجًا أطلقا عليه اسم تاجيت (Taggit) يُسْتَخْدَم في عمل ترميز لأنواع الكلمات داخل تلك الذخيرة البالغ عدد كلماتها مليون كلمة، واستخدما في ذلك عدد ٨١ علامة ترميز لتصنيف الكلمات، ووصل عدد القواعد التي لجئا إليها لتنفيذ تلك العملية حوالي ٣٣٠٠ قاعدة، ووصلت نسبة الدقة في تميز أنواع الكلمات إلى ٧٧%[29].

١ــ توزيع المادة اللغوية داخل ذخيرة براون[30]

تتوزع المادة اللغوية في ذخيرة براون على ١٥ فئة يُرْمَز لها بالحروف من A - R، من بينها الفئات من A - J تندرج تحت بند النصوص الإخبارية، أما الفئات من K-R فتندرج تحت بند الأعمال الخيالية الإبداعية، أما الأرقام الموجودة في كل فئة فتشير إلى عدد العينات في كل منها.

A) الجرائد والمجلات: في مجال الأخبار.

	جرائد يومية	مجلات أسبوعية	إجمالي
السياسية	١٠	٤	١٤
الرياضية	٥	٢	٧
المجتمع	٣	صفر	٣
تقارير صحفية	٧	٢	٩
التجارة	٣	١	٤
الثقافة	٥	٢	٧

B) الجرائد والمجلات: في مجال الاجتماعيات

	جرائد يومية	مجلات أسبوعية	إجمالي
قضايا اجتماعية	٧	٣	١٠
شئون خاصة	٧	٣	١٠
مراسلات القراء	٥	٢	٧

C) الجرائد والمجلات: في مجال النقد

(موضوعات النقد هي: المسرحيات، والكتب، والموسيقى، والرقص).

D) الديانات

كتب	٧
دوريات	٦
منشورات	٤

E) المهارات والهوايات

كتب	٢
دوريات	**34**

F) الحكايات الشائعة

كتب	٢٣
دوريات	٢٥

G) الأدب، والتراجم، والسير الذاتية

كتب	٣٨
دوريات	٣٧

H) متفرقات

مستندات حكومية	٢٤
تقارير مالية	٢
تقارير صناعية	٢
دليل الجامعات	١
الهيئات الصناعية	١

I) كتب تعليمية وعلمية

العلوم الطبيعية	١٢
الطب	٥
الرياضيات	٤
علم الاجتماع والسلوكيات	١٤
السياسة والقانون والتربية	١٨
علم الأجناس	١٨
الهندسة والتكنولوجيا	١٢

J) القصص العادية

قصص	٢٠
قصص قصيرة	٩

K) القصص البوليسية

قصص	٢٠
قصص قصيرة	٤

L) القصص الخيالية

قصص	٣
قصص قصيرة	٣

M) قصص المغامرات والرحلات

قصص	١٥
قصص قصيرة	١٤

N) القصص الرومانسية

قصص	٣
قصص قصيرة	٣

O) الفكاهة

قصص	٣
سرد	٣

وبعد تحديد أنواع النصوص وتفريعاتها الدقيقة وعدد العينات في كل نوع، تم جمع تلك العينات وإدخالها إلى ذاكرة الحاسب. وفي البداية كان هناك التزام بالأعداد المقررة لكل فرع في اختيار النصوص؛ وبحيث لا يقل كل نص عن ٢٠٠٠ كلمة، مع مراعاة أن تنتهي كل عينة بجملة كاملة. كما يتم وضع رقم كودي لكل نص عبارة عن رمز التصنيف الذي ينتمي إليه النص مضافاً إليه عدد مكون من رقمين.

٢ـ الإصدارات المختلفة لذخيرة براون

بعد إنشاء ذخيرة براون تتابعت أعمال تطويرها، وتم ذلك من خلال ستة إصدارات، وكان كل إصدار من هذه الإصدارت، يستهدف مجموعة من الباحثين في أغراض مختلفة، وسوف نتحدث عن تلك الإصدارات تفصيلاً فيما يلي:

أ) الإصدار "A":

كان هذا الإصدار هو الشكل الأول من الذخيرة اللغوية براون. وقد تم بناؤه في عام ١٩٦٤، إلا أن ذلك الإصدار قد تأثر بإمكانات الحاسب آنذاك، بالإضافة إلى استخدام تقنيات معقدة لعمليات التكويد.

ب) الإصدار "B":

وهو الإصدار الذي ظهر بعد إجراء معالجة على الإصدار"A"، وتمحورت تلك المعالجة حول حذف علامات الترقيم والكلمات المُشبَّكة والرموز المستخدمة في المعادلات. ولذلك يطلق على ذلك الإصدار اسم

النسخة المُجَرَّدة (Stripped)؛ حيث إن هذه المعالجـــة مفيـــدة فـــي عمـــل الدراسات الخاصة بالكلمات المفردة، هذا بالإضافة إلى أن هذه النـــسخة قـــد مَهَّدَت السبيل أمام كل من فرانسيز وكوسيرا لعمل الإحصائية الخاصة بمعدل تكرار الكلمات في الإنجليزية الأمْيركية.

ج) الإصدار "C":

وهو الإصدار الذي يحتوي ترميز الأنواع النحوية للكلمات، وتم تنفيذ هذا الإصدار بناءً على الإصدار"B". وقد تم توحيد الإشارات المعبـــرة عـــن نوع الكلمات، واستُخدِم في ذلك عدد ٨١ علامة للترميز النحوي.

د) الإصدار بِرْجِن الأول:

هذا الإصدار والإصدار التالي له قام بهما مركز الحاسب الآلي بجامعة بِرْجِن تحت إشراف العالم جوستين (Jostein)، وقد احتفظ هـــذا الإصـــدار بالحروف الكبيرة والحروف الصغيرة في الإنجليزية وترميز علامات الترقيم، مع الاستعانة بأقل قدر من الأكواد، بالإضافة إلـــى الاحتفـــاظ بالمعلومـــات التحريرية للنصوص.

هـ) الإصدار بِرْجِن الثاني:

الاختلاف الوحيد بين هذا الإصدار والإصدار السابق له هو تقليل قدر المعلومات التحريرية الخاصة بالنصوص؛ بالإضافة إلى وجود نظام متكامل يُمَكِّن من فهرسة الكلمات التي تحتويها الذخيرة.

و) الإصدار براون مارك:

أُنهي هذا الإصدار من خلال جامعة ستانفورد، وكان الغرض منه تمكين ذخيرة براون من التوافق مع برمجيتين شائعتين للاستعلام، وإحدى هاتين البرمجيتين تتيح إمكانية الاستعلام داخل الذخيرة عن كلمات محددة أو جمل كاملة طبقًا لسياق معين. أما الأخرى فتتيح إمكانية الاستعلام عن النصوص التي تندرج تحت سياق معين طبقًا لبعض الكلمات المفتاحية.

وإذا نظرنا إلى ذخيرة براون مقارنة بباقي الذخائر اللغوية لوجدنا أنها اتبعت تصميمًا في غاية الدقة من حيث العينات وأسلوب جمع المادة اللغوية.

ثالثًا: ذخيرة لوب

صاحِبُ اقتراح إنشاء هذه الذخيرة اللغوية هو العالم جيفري ليتش (Geoffrey Leech) من جامعة لانكاستر، وكان ذلك في سبعينيات القرن العشرين؛ ولكنها في النهاية نُفِّذَت بإشراف العالم ستيج جوهانسون (Stig Johansson) من جامعة أوسلو بالنرويج، وفي النهاية تم تحميلها باسم ذخيرة لوب **LOB** في مركز الحاسب والعلوم الإنسانية التابع لجامعة بِرْجِن [31].

وهذه الذخيرة بمثابة توأم لذخيرة براون، وكان الغرض من إنشائها دراسة اللغة الإنجليزية. وقد جاء تقسيم المواد داخل ذخيرة لوب مطابقًا تمامًا لتقسيمها في ذخيرة براون حتى يتسنى عمل الدراسات المقارنة بين النموذج الأميركي والبريطاني في اللغة الإنجليزية[32] ويوضح الجدول التالي توزيع المواد اللغوية داخل هاتين الذخيرتين.

جدول (٢-٢): الهيكل الداخلي لذخيرتي براون ولوب

الفئة	شكل المادة اللغوية	عدد العينات	
		ذخيرة براون	ذخيرة لوب
A	الجرائد والمجلات: في مجال التقارير الإخبارية	٤٤	٤٤
B	الجرائد والمجلات: في مجال الاجتماعيات	٢٧	٢٧
C	الجرائد والمجلات: في مجال النقد	١٧	١٧
D	الديانات	١٧	١٧
E	المهارات والهوايات	٣٦	٣٨
F	حكايات شائعة	٤٨	٤٤
G	الأدب	٧٥	٧٧
H	متفرقات	٣٠	٣٠
J	التعليم	٨٠	٨٠
K	قصص عادية	٢٩	٢٩
L	قصص بوليسية	٢٤	٢٤
M	قصص خيال علمي	٦	٦
N	مغامرات	٢٩	٢٩
P	قصص رومانسية	٢٩	٢٩
R	فكاهة	٩	٩
الإجمالي		٥٠٠	٥٠٠

كما صمم فريق العمل البحثي في جامعة لانكاستر مجموعـــة رمـــوز مكوَّنَة من ١٣٣ علامة. واستخدموا أسلوبًا خاصًّا لترميـــز أنـــواع الكلمـــات مختلفًا عن ذلك المستخدم في ذخيرة براون لتمييز الأنواع النحوية للكلمـــات

داخل ذخيرة LOB؛ حيث طوروا برمجية لترميز أنواع الكلمات أطلقوا عليها اسم تاجيت (Taggit). وقد اعتمد هذا النظام على الكلمات التي تم ترميزها من حيث النوع النحوي للكلمة في ذخيرة براون، وعن طريق قوانين الإحصاء تمكّنت البرمجية من وضع مصفوفة تُعَبِّر عن احتمالات التغير في نوع الكلمة بناءً على مصاحباتها من الكلمات الأخرى داخل النص، وعن طريق المعلومات الإحصائية تمكّنت هذه البرمجية من وضع ترميز لأنواع الكلمات داخل ذخيرة لوب LOB بأسلوب أكثر ذكاءً. وقد تَمكّن فريق العمل من خلال تلك البرمجية من الارتقاء بنسبة الدقة في تحديد أنواع الكلمات داخل الذخيرة البالغ عدد كلماتها مليون كلمة إلى نسبة ٩٦-٩٧%. وتُعتَبَر هذه الطريقة في التعامل مع الذخائر اللغوية إنجازًا عظيمًا تم التوصل إليه في حقل معالجة اللغات الطبيعية.

كما اتضح للعلماء العاملين في مجال معالجة المعلومات اللغوية من خلال تلك الحقيقة العلمية مدى محدودية وسائل الذكاء الاصطناعي، التي تعتمد على الحقائق المجردة مقارنة بتلك القوة الهائلة والأفق البغيد الذي تطرحه الأنظمة التي تعتمد على الإحصاء والاستقراء في التوصل إلى سبر أغوار الحقائق اللغوية داخل النص. ولعلنا لا نبالغ عندما نقول: إن برنامج (Taggit) قد فتح مجالا وأُفُقًا واسعًا في حقل معالجة اللغات الطبيعية، الأمر الذي أعطى إشارة الانطلاق لتطوير منهجيات التعامل مع الذخائر اللغوية خلال تسعينيات القرن العشرين. واعتمادًا على الأساس الذي قدمته ذخيرة لوب LOB في ترميز أنواع الكلمات نحويًا، أعلن كلٌّ من جوهانسون

(Johansson) وهوفلاند (Hofland) النتيجة الإحصائية لمعدل تكرار الكلمات وأنواعها النحوية في اللغة الإنجليزية عام ١٩٨٩ [33]، كما قام فريق من الباحثين بجامعتي لانكاستر وليدز بعمل ترميز للقواعد النحوية داخل ذخيرة لوب؛ تمهيدًا لاستخدام ذلك الترميز في عمل تحليل نحوي آلي قائم على نماذج علم الاحتمالات.

رابعًا: ذخيرة إل إل سي LLC الشفهية

كان إنشاء الذخائر اللغوية الثلاث السابقة بمثابة وضع نهاية لعصر إهدار طاقة الإنسان ووقته في جمع المادة اللغوية، وإرساء لمكانة علم الذخائر اللغوية ليحتل موقعه المناسب في الدراسات اللغوية. إن ظهور الذخائر اللغوية الثلاث قد مَكَّنَ العلماء من تطوير أسلوب جمع المواد اللغوية من مستوى الكلمة والعبارة والجملة البسيطة كما كان يحدث في المراحل الأولى من تاريخ جمع المواد اللغوية بالطرق اليدوية لينتقل مباشرةً إلى مستوى جمع النصوص اللغوية الكاملة، كما ساهم أيضًا في تطوير نطاق المادة التي يتم جمعها من إطار مُحَدَّد إلى أُطُر وأفرع اللغة المختلفة. إلا أن تلك الذخائر الثلاث كانت تُرَكِّز على المستوى التحريري من اللغة وخاصة الذخيرتين الأخيرتين، ولم يكن هناك مجال لجمع مواد لغوية شفهية. ولذلك بدأ العمل في إنشاء ذخيرة للنصوص الشفهية عام ١٩٧٥.

في ستينيات القرن العشرين، أجرى العالم الشهير كويرك من جامعة لندن دراسة عن أساليب استخدام اللغة الإنجليزية سجل خلالها مواد شفهية

وصل حجمها إلى ما يزيد عن ألفي ساعة من حوارات ومواد إذاعية وغير ذلك، وبعد ذلك حوّل تلك المادة اللغوية الشفهية إلى صورة ورقية. وفيما بعد تولى العالم سڤارتڤيك (Svartvik) من جامعة لوند السويسرية مهمة تحويل تلك المادة إلى نسخة رقمية يتم التعامل معها من خلال الحاسب الآلي. وفي الوقت ذاته أطلق العلماء مشروع ملخص اللغة الإنجليزية الشفهية **The Survey of Spoken English** الذي أُطلِقَ عليه اختصارًا اسم إس إس إي **SSE**[34] والذي يُعْتَبَر في حقيقة الأمر توأم مشروع إس إي يو **SEU** الذي أشرنا إليه سابقًا. وكان الغرض من هذا المشروع استغلال قدرات الحاسب العالية في معالجة البيانات آليًا للتوصل إلى المعلومات الأولية عن اللغة الشفهية الإنجليزية من متن تلك الذخيرة. وقد شملت عملية ترميز الذخيرة إجراء تحليل للإيقاع ووحدات التنغيم، وأصوات التوكيد والأساليب المختلفة للغة الشفهية، الأمر الذي اعتُبِرَ مادة أولية ذات قيمة عالية لدراسة اللغة الإنجليزية الشفهية. وقد اكتمل العمل في مشروع ذخيرة إس إس إي **SSE** عام ١٩٨١. وقد أُطلِق على هذه الذخيرة الشفهية اسم إل إل سي **LLC**، وهو اختصار لـ **London-Lund Corpus of Spoken English** بمعنى ذخيرة لندن – لوند للغة الإنجليزية الشفهية.

وقد بدأت ذخيرة **LLC** بعدد من النصوص يصل إلى ٨٧ نصًا ويحتوي كل نص حوالي ٥٠٠٠ رمز. ولتسهيل عملية البحث داخل تلك الذخيرة قام المصممون بعمل تصنيف دقيق وفهرسة جيدة لمحتوياتها، وقد قُسِّمَت النصوص إلى خمسة تصنيفات رئيسة؛ هي:

١- حوارات وجهًا لوجه.

٢- حوارات تليفونية.

٣- مناقشات، وتحقيقات صحفية، ومجادلات.

٤- مناقشات جماهيرية حية، ومناظرات، ومحاضرات دون إعداد سابق.

٥- محاضرات جماهيرية مُعَدَّة سابقًا.

وبعد ذلك، تم عمل تصنيفات أكثر دقة في إطار تلك التصنيفات العامة، ثم أُعطي لكل صنف من هذه النصوص رقم مضافًا إليه الحرف S. وبالإضافة إلى ما قام به سفارتفيك (Svarvtik) من عمل ترميز للهجات والإيقاع داخل مقاطع الكلام في الذخيرة، فقد وضع تصميمًا دقيقًا لحزمة من برمجيات الفهرسة أُطْلِقَ عليها اسم "الكلمات المفتاحية داخل السياق Key Word in Context" ويُطْلَقْ عليها اختصارًا KWIC. وهذه الحزمة البرمجية لا تُسَهِّل عمليات فهرسة النصوص فحسب، بل يمكن الاستعانة بها في البحث عن مقاطع ذات خصائص لغوية معينة داخل النصوص، لدرجة أنه أصبح من الممكن التعرف على عدد مرات ظهور الكلمات ونوعها النحوي وعلاقتها الاقترانية مع غيرها من الكلمات داخل أي نص. وهذا المستوى من الفهرسة لا يتطلب عمل تكويد لكل فقرة من النص فقط، وإنما يصل الأمر إلى تصميم حزمة من الرموز تكفي لكل أنواع الكلمات التي تحمل خصائص النص. وفي أثناء عملية ترميز الأنواع النحوية للكلمات تكون البداية بكتابة حرف إنجليزي كبير يشير إلى النوع النحوي للكلمة، وبعد ذلك يضاف إلى

هذا الحرف علامة أخرى للتعبير عن أشكال التغير المختلفة التـــي تحــدث للكلمة. على سبيل المثال، يضاف إلى الحرف N العدد ٢ في صورة (N+2) للتعبير عن صيغة المثنى من الاسم، وتُستَخْدَم مثلاً الصيغة (N+z) للتعبيـــر عن الحالات الإعرابية المختلفة التي تلحق بالاسم وغير ذلك. وبالإضافة إلى ذلك، ومن أجل عمل دراسات أكثر دقة على قواعد اللغة الــشفهية، قـــاموا بتصميم مجموعة من العلامات النحوية للتعرف على وحدات التحليل النحوي. فضلاً عن قيامهم بتصميم برنامج لتحليل التعبيرات اللغوية. وقد وصل حجـــم ذخيرة إل إل سي LLC اللغوية في النهاية إلى نصف مليون كلمة.

خامسًا: ذخيرة كوبويلد Cobuild

الاسم الكامل لذخيرة كوبويلد (Cobuild) هو (**Collins Bermignhan University International Language Database**)، وهي ذخيـــرة للغـــة الإنجليزية تم تأسيسها بالتعاون بين دار نشر جامعـــة كـــولينز الإنجليزيـــة وجامعة برمنجهان. ويتمثل الغرض من بناء هذه الذخيرة اللغوية في إجـــراء دراسات معجمية على أساس المواد اللغوية المخزنة بداخلها. وقد تم إنـــشاء ذخيرة كوبويلد بدعم وتشجيع من العالم جون سِنكلير (**John Sinclair**) فـــي ثمانينيات القرن العشرين، وعلى أساس ذلك ابتكـــر حزمـــة مـــن الأدوات البرمجية وقام بتطويرها لإدارة المادة اللغوية المخزنة بها وتحليلها، واستطاع أن يُكَوِّن فريقًا من الخبراء المتخصصين في علم المعـــاجم وعلـــم الـــذخائر اللغوية[35]. وقد كانت المبادئ التي وُضِعَت لاختيار المواد اللغويـــة لـــذخيرة كوبويلد في الثمانينيات كما يلي:

١- اللغة التحريرية تمثل نسبة ٧٥%، والشفهية نسبة ٢٥%.

٢- يُشْتَرَط في المواد اللغوية التي يتم جمعها أن تكون من اللغة الإنجليزية القياسية، ولا يتم الجمع من اللغة الشفهية غير القياسية. وتمثل المادة المجموعة من الإنجليزية البريطانية نسبة ٧٠%، أما الإنجليزية الأميركية فتمثل ٢٥%، أما باقي مناطق اللغة الإنجليزية فقد تم تمثيلها بنسبة ٥% من المادة اللغوية.

٣- تعكس المادة اللغوية استخدامات اللغة الإنجليزية المعاصرة، مع السعي قدر الإمكان لأن تكون المادة اللغوية حديثة.

٤- لا يتم جمع الشعر أو المسرح أو النصوص التقنية.

٥- تتمثل مصادر المادة اللغوية في البالغين من الأعمار فوق سن ١٦ سنة، ولا تقل نسبة الأعمال النسائية عن ٢٥% من إجمالي الذخيرة اللغوية.

٦- المادة المجموعة لا تكون في صورة عينات أو مقتطفات، وإنما في صورة كاملة أو أجزاء كبيرة من نصوص، في حدود سبعين ألف كلمة، ليتناسب ذلك مع الدراسات القائمة على مستوى النصوص الكاملة.

وصل حجم المادة الأولية التي تم جمعها في البداية إلى عشرين مليون كلمة، ويرجع الفضل إلى هذه الذخيرة اللغوية الكبيرة في إصدار (معجم كوبويلد للغة الإنجليزية)[36] **COBUILD English Language Dictionary** الذي أصدرته دار نشر كولينز عام ١٩٨٧ الأمر الذي اعتُبِرَ أول حدث من

نوعه في مجال تأليف المعاجم على مستوى العالم. ويتمثل الاختلاف بين معجم كوبويلد وغيره من المعاجم في أنه يحتوي أمثلة لغوية دقيقة وواضحة عن الكلمات وشرحها وأساليب استخدامها، وأن جميع الأمثلة المذكورة مصدرها لغة حقيقية، ولم تتم صياغتها بمعرفة مؤلفي المعجم. وتُعْتَبَر تجربة مؤلفي هذا المعجم مثالاً نموذجيًا لإمكانية الاعتماد على الذخائر اللغوية ومنهجية تحليل المواد اللغوية فيها لإجراء دراسات لغوية واسعة النطاق.

وتُستَخْدَم ذخيرة كوبويلد اللغوية بصورة أساسية في الدراسات المتعلقة بالمفردات، والمعاني، والقواعد، بالإضافة إلى طرق الاستخدام اللغوي في اللغة الإنجليزية. كما يتم إمدادها بالمواد اللغوية الحديثة بصورة متتالية الأمر الذي جعل من ذخيرة كوبويلد اللغوية ذخيرة ديناميكية تتغير بصورة متوالية.

وقد اكتشف العلماء من خلال دراساتهم لذخيرة كوبويلد على مدى عشر سنوات أن إطار المفردات في اللغة الإنجليزية متسع للغاية، بالإضافة إلى أن تلك المفردات يتم استخدامها بطرق متعددة. وذلك يُثْبِت أن الدراسات اللغوية يلزمها من العينات ما يتناسب مع هذا الحجم من الاستخدام المتنوع للغة ومفردتها. ويطلق على هذه الذخيرة في الوقت الحالي اسم "بنك اللغة الإنجليزية" (The Bank of English) حيث وصل حجم المادة اللغوية التي تضمها ٣٢٠ مليون كلمة، وقد أُجْرِيَت عملية ترميز لأنواع الكلمات داخل هذه الذخيرة اللغوية، بالإضافة إلى عمل تحليل نحوي جزئي للمادة اللغوية بها يصل إلى ٢٠٠ مليون كلمة. والمادة اللغوية المتضمَّنة في تلك الذخيرة كلها مواد حديثة؛ حيث إن معظم النصوص الموجودة بها من النصوص التي

ظهرت بعد عام ١٩٩٠، وتشمل المواد التحريرية بها نصوصًا قصصية، وغير قصصية، بالإضافة إلى الجرائد والمجلات والنشرات والمراسلات والتقارير وغير ذلك. أما المواد الصوتية فتشمل حوارات يومية، وإذاعة، ومؤتمرات ومقابلات، ومناقشات... إلخ. ويرى سنكلير أن هذه الذخيرة اللغوية تُقَدِّم أمثلة حقيقية وموضوعية عن اللغة الإنجليزية في مجالات الحياة اليومية لغالبية الناس وذلك في مستويات الاستماع والحديث والقراءة والكتابة.

كما يقدم كوبويلد برمجية تتيح لمؤلفي المعاجم وعلماء اللغة إجراء تحليل معقد على المادة اللغوية داخل الذخيرة؛ حيث يمكن لهذه البرمجية أن تقوم بالمهارات التالية:

١- عمل مسح لأنماط التراكيب التي تأتي مع كلمة محددة واستخراجها.

٢- الكشف عن معدل تكرار الكلمات.

٣- استخراج أمثلة واقعية عن الاستخدام اللغوي لكلمة من الكلمات، بل إجراء تحليل لتلك الأمثلة، ونسخ النتائج على وسائط التخزين الصلبة.

في عصر تكنولوجيا المعلومات، أصبحت هناك حاجة مُلِحَّة تتزايد بصورة مستمرة لأن يقوم الحاسب بعمل معالجة المواد اللغوية، بما في ذلك معالجة الشكل الكتابي للغة، أو الاستكشاف والإبحار داخل المعلومات اللغوية، أو الترجمة الآلية وغير ذلك. وتُعْتَبَر خدمات المعلومات الرقمية من

الأجزاء التي تلعب دورًا محوريًا في هذا المجال. ويمكن لذخيرة كوبويلد أن تقدم معلومات غزيرة فيما يتعلق بالمفردات والقواعد، الأمر الذي يضمن تطوير التكنولوجيا الخاصة بمعالجة اللغة في المجالات سالفة الذكر وتحسينها.

سادسًا: ذخيرة لونجمان اللغوية Longman

ظهرت هذه الذخيرة نتيجة العمل الذي قامت به لجنة ذخيرة لونجمان Longman Corpus Committee في الفترة من يناير ١٩٨٨ إلى نوفمبر ١٩٩٠ وقد أشارت سومرز (Summers) في الوثائق الخاصة بإنشاء الذخيرة إلى الخصائص التي اتبِعَت لتصميم ذخيرة لونجمان وهي كالتالي:

١ـ الهدف من إنشاء ذخيرة لونجمان هو بناء ذخيرة لغوية متعددة الأغراض بمعايير موضوعية

كان الهدف من ذخيرة لونجمان هو إنشاء ذخيرة لغوية جديدة تمامًا للغة الإنجليزية، وذلك عن طريق جمع كميات كبيرة من النصوص طبقًا لمقاييس ومعايير مناسبة وواضحة؛ وذلك بغرض وضع تلك الذخيرة لخدمة صناعة المعاجم وخدمة المحافل العلمية. في البدايات الأولى للذخائر اللغوية تكونت ذخيرتا براون وإس إي يو (SEU) طبقًا لإطار معين تم وضعه سابقًا لتحديد حجم المادة اللغوية. وكان الأسلوب المُتَّبَع دائمًا هو أن المادة اللغوية يتم جمعها بصورة مباشرة (فيما عدا النزر القليل) ولم يعتمد مصممو هاتين الذخيرتين على أسلوب المواد القياسية التي يتم تحديدها سابقًا في جمع المادة اللغوية، الأمر الذي تسبب في وجود أمثلة لغوية مُشوَّهة. أما ذخيرة لونجمان فقد اتَّبَعَت منهجية مختلفة عن الذخائر السابقة في بناء الهيكل الأساسي لها.

٢ــ مبادئ تصميم ذخيرة لونجمان

أ) الاعتماد على جمع النصوص التي تنتمي إلى أهل اللغة الإنجليزية الأصليين

حيث اعتمد مصممو ذخيرة لونجمان بصورة كبيرة على تراكم المعرفة اللغوية لأهل اللغة الإنجليزية على مدى أزمان طويلة، وأعطوا أولوية كبيرة للْحِس اللغوي لأهل اللغة الإنجليزية. واعتمادًا على ذلك المبدأ تمكن مؤلفو المعاجم من تحليل المعلومات اللغوية الخام داخل الذخيرة اللغوية وتفسيرها، بالإضافة إلى أنهم تمكنوا من تمييز ما يمكن القياس عليه وما هو نادر الظهور داخل اللغة. ولكن في الوقت ذاته قدمت تلك الذخيرة لمؤلفي المعاجم كمية كبيرة من المعلومات تفوق إحساسهم الموضوعي تجاه اللغة بشكل كبير، واستطاعت الذخيرة في أغلب الأحيان أن تتغلب على سوء الفهم الذي كان يبدو على الناس في السابق تجاه بعض الكلمات وأنماط القواعد النحوية، بالإضافة إلى أنها ألقت الضوء على العديد من الخصائص اللغوية الجديدة للكلمات. وهذا ما لم يكن لمؤلفي المعاجم تصوره في الماضي، فالأهم عند بناء الذخيرة اللغوية هو احترام شعور أهل اللغة الأصليين تجاه اللغة، بالإضافة إلى سلطة الذخيرة اللغوية في إقرار ما يتعلق بذلك الشعور من خصائص اللغة.

ب) إتاحة الذخيرة اللغوية لخدمة البحث العلمي

كان الهدف هو إنشاء ذخيرة لغوية متوازنة تُعَبِّر عن اللغة الإنجليزية خلال القرن العشرين، وبحيث تُغَطي تلك الذخيرة كلا من النموذج الأميركي والبريطاني من اللغة الإنجليزية. كما تشمل التحولات الرئيسة على ساحة

القوميات الأخرى الناطقة باللغة الإنجليزية، ويكــون ذلــك مُــشتَمِلا علـــى الإطارين التحريري والشفهي من اللغة.

الاستخدام الأساسي للذخيرة هو: إتاحة بيانات لغوية موضوعية يمكن الاعتماد عليها في تقديم تصور إجمالي للغة يمكن الاعتماد عليه، بالإضـــافة إلى وضع الأساس لتأليف المعاجم وكتب النحو وجميع المؤلفات الأخرى في علوم اللغة.

الاستخدام الفرعي للذخيرة هو: تقديم ذخيرة متوازنة الحجم عن اللغـــة الإنجليزية في القرن العشرين.

ج) البدء بتطوير الجزء المتعلق بالمستوى التحريري.

٣- منهجية اختيار المادة اللغوية

المادة اللغوية موضع الاختيار ينبغي أن تعبر عن اللغة الإنجليزية في القرن العشرين ابتداءً من عام ١٩٠٠، مع التركيز بصورة أكبر على المواد اللغوية المعاصرة. وتُقَسَّم النصوص من حيث النوع إلى نصوص معلوماتية **(Informative)** ونصوص إبداعية **(Imaginative)** ويشغل كل مــن هــذين النوعين نسبة ٦٠% و٤٠% على التوالي، إلا أن القائمين على بناء ذخيـــرة لونجمان يعتقدون أن النصوص القصصية أكثر تأثيرًا من النــصوص غيـــر الأدبية، بل لها عدد أكبر من القراء، ويمكن إثبات ذلك ببساطة عن طريـــق الاطلاع على المعلومات الإحصائية عن حالات الاستعارة والقــراءة داخـــل المكتبات.

أ) موضوعات النصوص: النصوص الحوارية أكثر مــن النــصوص الإبداعية

تتوزع المادة اللغوية داخل ذخيرة لونجمان على عشرة مجالات. مــن بين هذه المجالات نصوص قصصية تصل إلى ٤٠% من حجــم الــذخيرة وتشمل هذه النسبة الشعر والمسرح والكوميديا. وتُعتَبَر النصوص التحريرية هي المحرك الأساسي لاختيار الموضوعات وليس الشكل الأدبي. ونسبة مــا تمثله الموضوعات العشر من حجم الذخيرة كما يلي:

١) الطبيعة والعلوم البحتة	٦ %
٢) العلوم التطبيقية	٤,٣ %
٣) العلوم الاجتماعية	١٤,١ %
٤) الشئون الدولية	١٠,٤%
٥) المال والتجارة	٤,٤ %
٦) الفنون	٧,٩ %
٧) العقيدة والفكر	٤,٧ %
٨) الترفية	٥,٧ %
٩) القصص	٤٠ %
١٠) الشعر والمسرح والكوميديا	٢,٣ %
الإجمالي	٩٩,٨%

ب) الخصائص الرئيسة للمستندات

من أجل تقسيم النصوص الموجودة داخل الــذخيرة اللغويــة حــسب الأنواع المختلفة للمستندات تم اتباع أربعة تقسيمات خارجية فرعية تشير إلى الخصائص الرئيسة للنص؛ مثل: المنطقة التي ينتمي إليها الــنص، وزمــن صدوره، ووسيط نشر النص، ورتبته. وجميع هذه الخصائص فيما عدا رتبة

النص بمثابة عنصر قياس موضوعي للنص، وكل نص من النصوص التحريرية ينبغي أن يتضمن إشارة إلى تلك الخصائص.

المنطقة التي ينتمي إليها النص: تعرضت مواد الذخيرة لإنجلترا بصورة أساسية. على أساس النسبة التالية: إنجلترا ٥٠%، وأميركا ٤٠%، والدول الأخرى ١٠%.

زمن صدور النص: لم تلجأ ذخيرة لونجمان إلى فكرة وجود فترة زمنية محددة لاختيار المادة اللغوية، وإنما احترمت منهجية الذخيرة التاريخية؛ حيث غطت المواد الإنجليزية منذ عام ١٩٠٠؛ ولذلك نجدها ملائمة بصورة أكثر لتغطية أهداف الذخائر اللغوية ذات الاستخدام العام. وتتضح نسبة توزيع النصوص زمنيًا من خلال الجدول التالي:

جدول (٢-٣): الفترات الزمنية للنصوص في ذخيرة لونجمان

التاريخ	النصوص الإبداعية	النصوص المعلوماتية
١٩٠٠ - ١٩٤٩	٣٠%	١٠%
١٩٥٠ - ١٩٦٩	٣٠%	٢٠%
١٩٧٠ - حتى الآن	٤٠%	٧٠%

وسيط النشر: مصادر جمع النصوص التحريرية تشمل الكتب، والجرائد، ومتفرقات. وهذه المتفرقات تشمل النشرات غير المطبوعة والإعلانات والتقارير التجارية والإخطارات الحكومية والمنشورات وغير ذلك. ومن بين النصوص التحريرية كان المصدر الأساسي للجزء الخاص بالنصوص الإبداعية هو الكتب، أما النصوص المعلوماتية فتشمل الكتب

والجرائد والمتفرقات. أما نسب الوسائط المختلفة فهي ٨٠% للكتب، و١٣,٣% للجرائد، و٦,٧% للمتفرقات.

رتبة النص: وهذه أصعب خاصية يمكن تحقيقها لضمان توافق الأنواع داخل الذخيرة؛ حيث تحتل النصوص الإبداعية بما تشمله من قصص أدبية الرتبة الأعلى بين النصوص، يليها النصوص الخاصة بالمعلومات والنصوص التكنولوجية. وبالإضافة إلى ذلك فقد صُمِّمَت ذخيرة لونجمان لتشمل بعض الخصائص الثانوية للنصوص.

انطلاقًا من المقاييس والقواعد السابقة، فقد انتقت ذخيرة لونجمان نصوصها من أصل ٢٠٠٠ نوع من المصادر اللغوية، من بينها أكثر من نص تم اجتزاؤه من كتب، وقد وصل حجم المادة اللغوية في الذخيرة بأكملها إلى ٢٨ مليون كلمة متاحة للباحثين من أجل الاستخدام في البحث اللغوي.

سابعًا: الذخيرة القومية البريطانية BNC

مما لا شك فيه أن الذخيرة القومية البريطانية التي أُنشِئَت في الفترة من عام ١٩٩١ حتى عام ١٩٩٥ تتسم من حيث تصميمها وأسلوب جمعها أنها أكبر الذخائر حجمًا على مستوى العالم. وقد شاركت الحكومة البريطانية بـ ٥٠% من قيمة تمويل إنشاء هذه الذخيرة، وقد تم تطوير هذه الذخيرة بالتعاون بين كل من دار نشر جامعة أكسفورد ومجموعة لونجمان ودار نشر تشامبرز والمكتبة القومية البريطانية وجامعة أكسفورد وجامعة لانكاستر. وقد قدم كل من هذه الهيئات جميع ما لديها من خبرات وإمكانات إدارة ونشر النصوص الإلكترونية وتأليف المعاجم ومجالات تحليل الذخائر اللغوية حتى يُمكِن العمل في إطار مشترك من أجل تصميم تلك الذخيرة اللغوية وتطويرها وترميزها. ونظرًا إلى أن ذخيرة بي إن سي BNC تشتمل على نصوص

تحريرية وأخرى شفهية بالإضافة إلى كبر حجم مادتها اللغوية، الأمر الذي ولّدَ الأمل أن تُصبحَ المصدر القومي الأساسي للاطلاع على المصادر الإنجليزية مثلها في ذلك مثل ذخائر إس إي سي SEC، وبراون، ولوب LOB وما قدمته من إسهام في حقل الدراسات اللغوية المبنية على ذخائر لغوية في أثناء عصر الجيل الأول من الذخائر. وقد كان الفكر التصميمي لذخيرة بي إن سي BNC يتمتع بقدر عالٍ من التوازن؛ حيث جمعت الذخيرة قطاعًا عريضًا من النصوص التحريرية والشفهية، الأمر الذي جعلها المصدر الأساسي الأكثر انتشارًا واستخدامًا في التطبيقات اللغوية التجارية والعلمية والتعليمية. تشمل ذخيرة بي إن سيBNC عدد ٤١٢٤ نص، تمثل اللغة التحريرية منها نسبة ٩٠%، أما اللغة الشفهية فتمثل ١٠%. وعلى الرغم من أن الجزء الشفهي الذي يصل إلى حوالي ١٠ مليون كلمة، يُعْتَبَر في الوقت ذاته أكبر نص شفهي يمثل اللغة الإنجليزية حتى الآن، فإن ليتش (Leech) قد لاحظ عام ١٩٩٣ أن ذخيرة بي إن سي BNC لم تستطع أن تحقق التوازن بين حجم المحتوى التحريري والمحتوى الشفهي في المادة اللغوية الممثلة بداخلها. ويقدم المصدر رقم[38] وهو موقع الذخيرة على شبكة المعلومات تعريفًا عن توزيع المادة اللغوية داخل ذخيرة بي إن سي BNC.

١ـ ذخيرة BNC للنصوص التحريرية

تضم ذخيرة بي إن سي BNC للنصوص التحريرية عدد ٣٢٠٩ نصوص، وكل نص من هذه النصوص يتمتع بالخصائص اللغوية للفئة اللغوية التي يندرج تحتها. وعند اختيار المادة اللغوية يتم التحقق من وجود ثلاث خصائص هي تاريخ الإصدار ووسيط النشر والمجال الرئيس الذي تندرج تحته تلك المادة.

أ) تاريخ الإصدار

جدول (٢-٤): تواريخ إصدار النصوص في ذخيرة بي إن سي BNC

التاريخ	عدد النصوص	نسبة التمثيل في الذخيرة
١٩٦٠-١٩٧٤	٥٣	١,٦٥%
١٩٧٥-١٩٩٣	٢٥٩٦	٨٠,٨٩%
من دون تاريخ	٥٦٠	١٧,٤٥%

ب) وسيط النشر

لا يتعدى حجم أي نص في ذخيرة بي إن سي **BNC** مهما بلغ حجمه عدد ٤٠ ألف كلمة، والجدول التالي يبين لنا نسبة المصادر التحريرية المختلفة وما يمثلها من وسائط نشر.

جدول ٢-٥: نسب توزيع الوسائط المختلفة من النصوص على ذخيرة بي إن سي

الوسيط	عدد العينات	نسبة التمثيل داخل الذخيرة
الكتب	١٤٨٨	٤٣,٣٦%
الجرائد	١١٦٧	٣٦,٣٦%
أوراق متفرقة (إعلانات ونشرات)	١٨١	٥,٦٤ %
أوراق متفرقة غير علنية (مراسلات ومذكرات)	٢٤٥	٧,٦٣ %
لغة شفهية مكتوبة في شكل تحريري	٤٩	١,٥٢ %

ج) المجال

هناك نسبة ٢٠% من النصوص داخل ذخيرة بي إن سي **BNC** تندرج تحت بند النصوص الإبداعية، ونُشِرَت جميعًا بعد عام ١٩٦٠، وهناك نسبة ٨٠% من النصوص تندرج تحت فئة النصوص المعلوماتية نُشِرَت جميعها

بعد عام ١٩٧٥. وتزيد نسبة عينات النصوص الإبداعية عن نسبة القصص التي تم جمعها من المجلات الأسبوعية؛ وذلك بسبب أن مصممي الذخيرة قد راعوا التأثير الثقافي المستمر للأعمال الأدبية في المجتمع، والجدول التالي يوضح ذلك بالتفصيل.

جدول (٢-٦): نسب توزيع المجالات داخل ذخيرة بي إن سي

المجال	عدد النصوص	نسبة التمثيل داخل الذخيرة
النصوص الإبداعية	٦٢٥	%١٩,٤٧
العلوم الطبيعية	١٤٤	% ٤,٤٨
العلوم التطبيقية	٣٦٤	%١١,٣٤
العلوم الاجتماعية	٥١٠	%١٥,٨٩
الشئون الدولية	٤٥٣	%١٤,١١
التجارة والمال	٢٨٩	% ٨,٨٥
الفنون	٢٥٩	% ٨,٠٧
العقيدة والفكر	١٤٦	%٤,٥٤
الترفيه	٣٧٤	%١١,٦٥
من دون تصنيف	٥٠	%١,٥٥

٢- ذخيرة بي إن سي BNC للنصوص الشفهية

تضم ذخيرة بي إن سيBNC مواد شفهية تصل إلى ١٠ ملايين كلمة، وقد أُخِذَت هذه المادة من مصدرين أساسين: مواد إدارية رسمية-(cotext-governed material) وعينات إحصائية. ويصل حجم المواد الإدارية الرسمية إلى ٦١٥٤٢٤٨ كلمة. أما العينات الإحصائية فيصل حجمها إلى ٣٨٤٥٧٥٢.

أ) النصوص الشفهية المأخوذة من المواد الإدارية الرسمية

من أجل الوصول إلى تغطية شاملة لأنواع النصوص الشفهية المختلفة ضَمَّت المواد الإدارية الرسمية الأنواع التالية: المحاضرات وما يتعلـق بهـا من أحداث تتصل بالمعلومات الخاصة بالعملية التعليمية؛ وتسجيلات الفيـديو واستطلاعات الرأي والاختبارات الشفهية وما شابه ذلك من أنشطة؛ بالإضافة إلى النشرات المصورة والخطب الحكومية والمؤتمرات العامة والاجتماعـات البرلمانية، وغير ذلك من الأعمال الحكومية والعامة؛ كما ضـمَّت الـذخيرة التعليقات الرياضية المُصَوَّرة وأنشطة النوادي والأحاديـث الإذاعيـة عبـر التليفون وغير ذلك من الأنشطة الترفيهية والحوارات. وقد تـم جمـع هـذه النصوص الشفهية بصورة منتظمة مـن عـدد ١٢ منطقـة علـى مـستوى بريطانيا.

والجدول التالي يوضح النسب التي يحتلها كل نوع من النصوص داخل الذخيرة.

جدول (٢-٧): توزيع المواد اللغوية الشفهية حسب السياق داخل ذخيرة بي إن سي

المواد الإدارية الرسمية	عدد النصوص	نسبة التمثيل داخل الذخيرة
مـواد تعليمية	١٤٤	%١٨,٨٩
شئـون عامة	١٣٦	%١٧,٨٤
شئـون إدارية	٢٤١	%٣١,٦٢
مـواد ترفيهية	١٨٧٧	%٢٤,٥٤
من دون تصنيف	٥٤	% ٧,٠٨

ب) النصوص الشفهية المأخوذة من العينات الإحصائية

المصدر الثاني للنصوص الشفهية يضم ١٢٤ نصًا قدمها عدد من المتطوعين، وقد وصل حجم هذه المادة إلى ٢٠٠٠ ساعة مُسَجَّلة تم تحويلها إلى مكافئ كتابي. وتضم الحوارات موضوع تلك التسجيلات جميع مناحي الحياة المختلفة، وتم جمعها من عدد ٣٨ منطقة إدارية على مستوى إنجلترا، وينتمي هؤلاء (ذكور وإناث) إلى ٤ طبقات اجتماعية مختلفة، وقد وُزِّعَت المادة متوازية على أشخاص من الذكور والإناث في أعمار تتراوح بين ١٥، و٦٠ عامًا أو أكثر. وكان كل متطوع أو متطوعة يحمل جهاز تسجيل محمولاً على الكتف، ويقوم بتسجيل حوارات كاملة بين الأشخاص على مدى يومين بصورة لا تلفت الانتباه. ثم يُخْطَر جميع الأشخاص أنه قد تم تسجيل حواراتهم، وإعطائهم سلطة حذف ما يرغبون من محتوى شريط الكاسيت. كما تم تسجيل جميع المعلومات الخاصة بمحيط الحديث، وجميع المعلومات المتعلقة بالمتحدثين. وتشتمل تلك التفاصيل الخلفية الثقافية للمتحدث، والحركات التي يقوم بها في أثناء الحديث، بالإضافة إلى مكان الحوار وزمنه وتاريخه والمستمعين ودرجة الارتجال وموضوع الحديث ونوع المشاركين وأعمارهم وجنسياتهم ووظائفهم ودرجة تعلمهم ودرجاتهم الاجتماعية وعلاقتهم بالمتحدث، ودرجة اللغة العامية المستخدمة وغير ذلك. وقد تم تحويل النصوص المُسَجَّلَة كافة إلى شكل تحريري كلمة بكلمة. مع عدم إغفال الوقفات، والتردد في أثناء الحديث، والأخطاء اللغوية، والتكرار سواء على مستوى الصوت أو الكلمة أو غير ذلك. بالإضافة إلى الحديث بصوت

عالٍ والتهامس وغير ذلك من الخصائص اللغوية الفائقة. ولـــم يـــتم وضـــع ترميز للخصائص الصوتية وتم الاكتفاء بوجود معلومات قليلة عن الإيقـــاع؛ ولذلك يمكن أن تستخدم ذخيرة بي إن سي BNC في إجراء ما نحتاج إليه من الدراسات الصوتية الدقيقة للغة.

٣- الخدمات التي تقدمها ذخيرة بي إن سي BNC

تم تحويل جميع النصوص الموجودة بذخيرة بي إن سي إلى نصوص مطابقة للمواصفات القياسية الدولية للنشر الإلكتروني (SGML) أما المـــادة اللغوية داخل الذخيرة فقد طُبِّق عليها نظام الترميز الخاص بالأنواع النحوية للكلمات الذي طورته جامعة لانكاستر والمعروف باسم CLAW. كما تُقَـــدِّم ذخيرة بي إن سي واجهة برمجية قوية للبحث عن المعلومـــات والاســـتعلام عنها داخل الذخيرة، وبإمكان تلك الواجهة البرمجية أن تُنَفِّذ عمليات البحـــث المعقدة والاستعلام داخل متن الذخيرة اللغوية.

ثامنًا: الذخيرة الدولية للغة الإنجليزية

في عام ١٩٨٨ طرح جرينباوم (Greenbaum) اقتراحًا بإنشاء ذخيرة دولية للغة الإنجليزية فائقة الحجم وتطويرها، وأطلق عليها اســـم الـــذخيرة الدولية للغة الإنجليزية (The International Corpus of English) ويطلـــق عليها اختصـارًا آي سي إي ICE. والهدف من تلك الـــذخيرة هـــو إجـــراء دراسات مقارنة بين اللغة الإنجليزية في عدد من الدول الناطقة بها. ويـــشمل نطاق المقارنة كلا من المستوى التحريري والشفهي للغة. وتـــضم الـــذخيرة الدولية للغة الإنجليزية الموجودة حاليًا عدد ٢٠ ذخيرة فرعية يـــصل حجـــم

المادة اللغوية في كل منها مليون كلمة. وقد تم اختيار المادة اللغوية فيها جميعًا من الأشخاص البالغين الذين تجاوزت أعمارهم ١٨ عامًا، الذين تلقوا تعليمهم الأساسي في المراحلة الثانوية وما بعدها. وتلك الذخيرة تضم الدول التي تتحدث اللغة الإنجليزية كلغة أولى؛ مثل: إنجلترا وأميركا وكندا والنمسا ونيوزيلندا وما إلى ذلك، كما تضم الدول التي تُعتَبَر اللغة الإنجليزية هي اللغة الرسمية لها أو لغة الغالبية فيها؛ مثل: الهند ونيجيريا وسنغافورة وغير ذلك. وتغطي المادة اللغوية التي جُمِعت لهذه الذخيرة الفترة من ١٩٩٠ حتى ١٩٩٣.

وعلى الرغم من أن الغرض من إنشاء الذخيرة الدولية للغة الإنجليزية كان إجراء دراسات مقارنة على اللغة الإنجليزية، فإن كل ذخيرة من الذخائر الفرعية المكوّنة لها يمكن الاعتماد عليها بمفردها في عمل دراسات وصفية للغة الإنجليزية في كل دولة بمفردها. وقد تمكن الباحثون من خلال هذه الذخيرة من اكتشاف بعض أنواع الاختلافات في استخدام اللغة الإنجليزية باختلاف الدول، على سبيل المثال التعرف على الدول التي تَستَخدِم عبارة (different from) والدول الأخرى التي تستخدم (different to)، كما يمكن اكتشاف نسب تواجد بعض الأساليب في عدد من المناطق دونًا عن غيرها مثل استخدام أسلوب النفي التوكيدي وغير ذلك.

كانت أول ذخيرة فرعية تم إنشاؤها في إطار الذخيرة الدولية للغة الإنجليزية هي الذخيرة الإنجليزية البريطانية. ومكونات هذه الذخيرة تتضح من خلال الجدول رقم . وبعد ذلك اتُّبِعَت باقي الذخائر اللغوية في الدول الأخرى الهيكلية نفسها في بناء ٢-٨ محتوي كل منها. وتحتوي كل ذخيرة فرعية في الذخيرة الدولية للغة الإنجليزية ٥٠٠ عينة لغوية تبلغ محتويات كل عينة ٢٠٠٠ كلمة تقريبًا، وتشغل المادة التحريرية نسبة ٤٠% من محتويات

الذخيرة، وغالبية المادة الشفهية التي تصل إلــى ٣٠٠ نــص عبــارة عــن حوارات عامة مُعلنة.

جدول (٢-٨): بنية الذخيرة الدولية للغة الإنجليزية

مستوى اللغة الشفهية (300 نص)

- **حوارات (180)**
  - **حوارات خاصة (100)**
    - حديث مباشر (90)
    - حديث عبر الهاتف (10)
  - **موضوعات علنية (80)**
    - محاضرات (20)
    - مناقشات عبر الإذاعة (20)
    - أحاديث إذاعية (10)
    - مناقشات مجلس الدولة (10)
    - مداولات قضائية (10)
    - شؤون تجارية (10)

إخبارية (علمية) (40)

أدب (10)

علوم اجتماعية (10)

علوم طبيعية (10)

تكنولوجيا (10)

إخبارية (عامة) (40)

حضارة (10)

علوم اجتماعية (10)

العلوم الطبيعية (10)

نصوص تقنية (10)

نصوص إخبارية (20)

تقارير (20)

التعليم (20)

نصوص إدارية / تعليم نظامي (10)

مهارات/ هوايات (10)

توجيهات (10)

افتتاحيات الأخبار (10)

خيال (20)

الفصل الخامس

بناء الذخائر اللغوية الصينية

إن الغرض من بناء ذخيرة للغة الصينية يبدأ أساسًا من الحاجة الماسة إلى التحليل الإحصائي للغة الصينية. ويشمل التحليل الإحصائي للغة الصينية الدراسات القائمة على معدل تكرار الرموز والمفردات الصينية؛ وذلك من أجل بناء قوائم الرموز والمفردات شائعة الاستخدام في اللغة الصينية. ويتطلب إجراء دراسات إحصائية على اللغة الصينية التعامل مع حقائق لغوية هائلة العدد، وهذا يعني بناء ذخيرة لغوية ذات حجم مناسب لتلك المهمة. في البدايات الأولى للذخائر اللغوية، كان هذا العمل الإحصائي يتم من خلال المجهود البشري. في الغرب، كان أول معجم لغوي قائم على دراسة إحصائية بالمجهود البشري أَلَّفَه العالِم اللغوي الألماني كايدن **(Kaeden)**(١) عام **1898**. أما في الصين، فكان أول إحصاء قائم على الاستخدام الحديث

---

(١) فريدريش ڤِيلْهِلْم كايدن **(Friedrich Wilhelm Kaeden)** (١٨٤٣-١٩٢٨): قام في عام ١٨٩٨ بالتعاون مع ثمانمائة شخص دراسة إحصائية يدوية على ذخيرة تحريرية كبيرة الحجم من أربعة عشر مصدرًا للجرائد والمجلات، وبمساعدة عدد خمسة آلاف ناسخ سريع على الآلة الكاتبة، وقد استغرق هذا العمل سبعة أعوام وصل خلالها حجم الكلمات التي أحصاها ما يقرب من أحد عشر مليون كلمة، من بينها عدد ٧٩٧١٦ كلمة وردت أكثر من أربع مرات في الذخيرة. وهذه النتيجة الإحصائية تم الاعتماد عليها في تأليف أول معجم تكراري للكلمات على مستوى العالم وهو ((المعجم التكراري للكلمات في اللغة الألمانية))، وقد اعتُبِرَ هذا العمل أول دراسة بحثية على الكلمات باستخدام الأسلوب الإحصائي بمفهومه الحديث. (المترجم)

للغة الصينية أجراه العالم التربوي تشن خه تشين في عام ١٩٢٨[40] بطريقة يدوية مع أربعة من مساعديه، وقد استغرق هذا العمل منهم ما بين عامين إلى ثلاثة أعوام، وقد استعانوا في ذلك بذخيرة لغوية مكونة من ٥٥٤٤٩٨ رمزٍ صيني، تندرج تحت ستة أنماط من النصوص. وقد أظهرت نتائج الإحصاء أن الرموز الصينية التي تكررت داخل هذه الذخيرة يصل عددها إلى ٤٢٦١ رمزًا، ومن بين هذه الرموز كان هناك أكثر من ٥٦٩ رمزًا تكرر أكثر من ثلاثمائة مرة، و١١٩٣ رمزًا تكرر أكثر من مائة مرة. وما زالت الأرقام الإحصائية التي تَوَصَّل إليها السيد تشن خه تشن ذات موثوقية علمية كبيرة حتى الآن. ومع حلول سبعينيات القرن العشرين، استعانت الصين بالجهود البشرية في إتمام عملية إحصاء لمعدلات تكرار الرموز الصينية في ذخيرة لغوية كبيرة الحجم تصل إلى أكثر من مليارِي رمز صيني، وهذا ما اشتُهِر باسم "المشروع رقم ٧٤٨".

ويُعْتَبَر التغلب على إشكاليات إدخال الرموز الصينية إلى الحاسب الآلي بمثابة وضع حجر الأساس بصورة ملموسة أمام أبحاث الذخائر اللغوية الرقمية باللغة الصينية. فمع نهاية سبعينيات القرن العشرين، أنشأت الصين على التوالي عددًا من الذخائر اللغوية الرقمية كبيرة الحجم التي تَستَخدِم منهجيات التحليل الكمي للغة الصينية، ويُعْتَبر نشر (المعجم الإحصائي لمعدل تكرار الرموز الصينية)، و(معجم الرموز الأكثر شيوعًا في اللغة الصينية) وغيرهما من المعاجم من أهم النتائج العلمية لهذه الفترة. ومع حلول ثمانينيات القرن العشرين، استعان العاملون في مجال معالجة المعلومات باللغة الصينية

بالذخائر اللغوية بصورة أساسية لعمل تمييز آلي لحـدود الكلمـات داخـل النصوص الصينية، وكان ذلك بمثابة المشروع الأساسي لبناء أنظمة الترجمة الآلية، والفهم الآلي للكلام، وغير ذلك من التطبيقات القائمة علـى معالجـة المعلومات باللغة الصينية. وبحلول تسعينيات القرن العشرين، وتمـشيًا مـع الارتفاع المتزايد لقدرة الحاسب الآلي على تخزين المعلومـات ومعالجتهـا، زادت أيضًا القدرة على معالجة الرموز الصينية. ففي البداية، كانت الذخائر اللغوية الصينية يتم بناؤها لأغراض ومهام بحثيـة محـددة، مثـل عمـل إحصاءات عن معدلات تكرار الرموز والمفردات الصينية، وهذا يختلف عن وظائف الذخائر اللغوية الحديثة التي يكون لها استخدامات واسعة النطـاق، التي لا تقتصر فقط على عمل الدراسات اللغوية، بل تمتد إلى عمل دراسات معالجة اللغات الطبيعية وما إلى ذلك. ويُعَرِّفُنا الفصل الحالي من الكتاب بعدد من الذخائر اللغوية القياسية للغة الصينية الحديثة.

أولاً: الذخائر المخصصة لإحصاء معدل تكرار الكلمات في اللغة الصينية

١- مشروع الذخيرة اللغوية العامة بجامعة اللغات والثقافة ببكين

لقد نتج عن هذا المشروع (معجم معدل تكرار الكلمـات فـي اللغـة الصينية الحديثة). ومن أجل تأليف هذا المعجم[41]، قام مركز الأبحاث اللغوية بجامعة اللغات والثقافة ببكين بإنشاء ذخيرة لغوية باللغة الـصينية الحديثـة تصل إلى مليوني رمز صيني. وفيما يلي نستعرض تفاصيل مَهَمَّة بناء هـذه الذخيرة: استخدام مختلف المواد اللغوية في مختلـف الموضـوعات لبنـاء الذخيرة، واستخدام أسلوب العمل اليدوي في عمل تمييز لحدود الكلمات، ثـم

الاستعانة بالحاسب الآلي في إنهاء مَهَمَّة إحصاء معدلات تكرار الرموز والمفردات الصينية داخل الذخيرة. وقد أتاح هذا المشروع إمكانية عمل دراسة استطلاعية شاملة عن الاستخدام الواقعي للرموز والمفردات الصينية في اللغة الصينية الحديثة، وقَدَّم دراسة إحصائية وتحليلية لظروف توزيع المفردات الصينية في مختلف النصوص بالإضافة إلى معدلات التكرار ونسب الاستخدام، فضلاً عن عمل مقارنة بين القيم التي تمت مشاهدتها وبين القيم المتوقَّعة؛ كما تم إحصاء معدلات ظهور الرموز الصينية في النصوص وتحليلها، ومقارنة قدرتها على تكوين الكلمات من خلال موقع كل منها داخل الكلمة. وكان الهدف من المشروع عرض ملامح استخدام الرموز والمفردات الصينية من خلال دراسات إحصائية، والتمييز بين درجات الاستخدام الأكثر شيوعًا لكل من الرموز والمفردات الصينية، وبعد إجراء اختبارات عشوائية لمعدلات التغطية، تم عمل تقييم لكامل النتائج التي تم التوصل إليها، وفي النهاية عُرِضَت قوائم المفردات والرموز النهائية وفقًا للشروط الموضوعة سابقًا مرفقًا بها المعلومات الإحصائية الخاصة بها.

ويقدم المرجع رقم [42] شرحًا تفصيليًا للأعمال التي قام بها فريق بناء تلك الذخيرة اللغوية.

أ) تحديد قواعد اختيار العينات اللغوية

في أثناء بناء الذخيرة اللغوية، استرشد مصممو الذخيرة بحدود اختيار المادة اللغوية في الذخائر السابقة، مع مراعاة مدى شيوع الاستخدام داخل المجتمع، وفي النهاية تم الاستقرار على أن يكون اختيار المادة اللغوية من

الجرائد والمناقشات السياسية، والمؤلفات العلمية، والحوارات اليومية، والمؤلفات الأدبية، وما إلى ذلك. وفي الوقت ذاته، تم التطرق إلى الاستعانة بنماذج من النصوص اللغوية في المرحلتين الابتدائية والإعدادية؛ بحيث تتمتع تلك النصوص بأسلوب لغوي قياسي، بالإضافة إلى مراعاة التدرج في المعلومات اللغوية والثقافية لمضمونها؛ ولذلك تم اعتماد النصوص اللغوية المعتَمَدَة من قِبَل الحكومة الصينية في الفترة ما بين ١٩٧٨-١٩٨٠ كمواد دراسية للمرحلتين الابتدائية والإعدادية. ويتمثل توزيع المواد اللغوية التي تم انتقاؤها للذخيرة المذكورة كما يلي:

المجموعة أ: سياسة، واقتصاد، وفلسفة، وقانون، وتاريخ، وجغرافيا، وشئون عسكرية وما إلى ذلك من المقالات الصحفية والمؤلفات التي تعكس ملامح الحياة في المجتمع الصيني الحديث، بما يزيد عن ٤٤٠ ألف رمز صيني وما يمثل نسبة ٢٤,٤% من إجمالي حجم الذخيرة.

المجموعة ب: معارف علمية، تم اختيارها من المستوى المتوسط لعلم الرياضيات، والأحياء، والطب، والهندسة، والتكنولوجيا، وعلوم الطيران والملاحة، والفضاء، وتاريخ العلوم، والسيرة الذاتية للعلماء، بالإضافة إلى مقالات علمية تتعلق بالملبس والمأكل والحياة فيما يصل إلى مائتين وتسعين ألف رمز صيني، وتشغل نسبة ١٥,٨% من حجم الذخيرة.

المجموعة ج: مواد شفهية حول مواقف للحياة اليومية؛ حيث تم اختيار أعمال مسرحية شهيرة تعكس مختلف جوانب الحياة (أعمال لكل من قوه مو روه، ولاو شه، وتيان خان، وتساو يو، وو تزي جوانغ)، بالإضافة إلى

حوارات كوميدية وحكايات قصصية وما إلى ذلك. هذا بالإضافة إلى التطرق إلى احتياجات الحياة اليومية من التعامل الشفهي اليومي، وقد تم الحصول على جزء من الذخيرة الشفهية عن طريق التسجيل في موضوعات محددة والتسجيل في موضوعات عشوائية لمادة شفهية. وقد وصل مجموع المادة اللغوية التي تم جمعها إلى ما يقرب من مائتي ألف رمز تحتل نسبة ١٠,٩% من حجم الذخيرة.

المجموعة د: روايات، وقصص قصيرة، وأعمال نثرية، وقصص أطفال تصل إلى ما يقرب من ٨٩٠ ألف رمز صيني، تحتل نسبة ٤٨,٧% من إجمالي حجم الذخيرة. وعند اختيار الأعمال الأدبية تمت مراعاة القواعد التالية:

١- التركيز على أن تكون الأولوية في الاختيار للأعمال الأدبية الصينية المتميزة منذ حركة الرابع من مايو، يلي ذلك الأعمال الأدبية التي تمثل الاستخدام اللغوي الجيد، مع تغطية معظم التيارات والأساليب اللغوية.

٢- الإكثار من اختيار الأعمال الأدبية التي تنتمي إلى الفترة ما بين أربعينيات إلى سبعينيات القرن العشرين. وبالنسبة إلى الموضوعات فقد تم التركيز على أن يتطرق الاختيار إلى معظم موضوعات الكتابة الأدبية (الحرب، والبناء، والمصانع، والقرية، والمدينة، والطبقات الاجتماعية، والأقليات القومية، والشخصيات التاريخية، ومختلف جوانب الحياة)، وقد حاول فريق العمل أن يحقق التوازن بين حجم المادة اللغوية في جميع المجالات.

٣- الاحتفاظ بتكامل النسخة الأصلية للعمل الأدبي؛ حيث رُوعي أن تحتوي الذخيرة نصوصًا قصيرة كاملة في إطار عشرة آلاف رمز صيني، أما غالبية الأعمال الأدبية الأخرى فقد رُوعي اختيار أجزاء كاملة منها بالإضافة إلى مراعاة أن تكون تلك الأجزاء مُمَثِّلة للعمل الأدبي.

وبالنسبة إلى مجمل المادة اللغوية، فقد كان الاهتمام بأن يتم اختيار المادة بأسلوب العينات المتساوية في الحجم فيما يتعلق بالأعمال التي تتناول المناقشات السياسية والمؤلفات العلمية؛ وذلك من أجل أن تكون المفردات ذات تغطية شاملة وموضوعية. أما فيما يتعلق بالأعمال المسرحية، فقد اهتم مصممو الذخيرة بعمل الإحصاء على مشاهد كاملة من المسرحيات؛ حيث اقتصر الأمر على الحوار المسرحي والحوارات الداخلية للشخوص المسرحية، ولم يتم إدراج وصف المَشَاهِد وغيرها من المواقف السرْدية داخل العمل المسرحي في عملية الإحصاء. أما فيما يتعلق بدروس الأدب واللغة في المرحلتين الابتدائية والإعدادية، فبالإضافة إلى النصوص الكاملة من اللغة الكلاسيكية، والشعر، والأعمال المترجمة عن اللغات الأجنبية، فقد تم عمل إحصاء بمعدل تكرار الرموز والمفردات داخل الأعمال بكاملها. مجمل القول: إن مصممي الذخيرة كانوا على قناعة أن أفضل قاعدة لاختيار عينات النصوص داخل الذخيرة هي مراعاة نسبة التمثيل وتعدد التخصصات، والتجانس، مع الاهتمام بالتعامل مع مشكلة الحجم الكلي للذخيرة بأسلوب علمي.

فحجم الذخيرة إذا كان صغيرًا جدًّا، لن تتمكن الذخيرة من التعبير عن المشكلة محل الدراسة، أما إذا كان الحجم كبيرًا فعلى الرغم من ارتفاع درجة دقة النتائج الإحصائية، فإن الفائض اللغوي في تلك الحالة سيكون كبيرًا جدًّا، فضلاً عن التكلفة الاقتصادية العالية. وقد راعى مؤلفو (معجم معدل تكرار الكلمات في اللغة الصينية الحديثة) مقدار التكامل بين هذين العنصرين بصورة كبيرة في عملية اختيار المادة اللغوية وتحديد نطاقها.

ب) النتائج البحثية لتلك الذخائر اللغوية

تختلف اللغة الصينية عن غيرها من اللغات الأجنبية التي تعتمد على الأبجدية الكتابية؛ حيث تفتقر اللغة الصينية إلى العلامات الصرفية التي تُعَبِّر عن نوع الكلمة، بالإضافة إلى عدم وجود علامات فاصلة بين الكلمات بعضها البعض، الأمر الذي يُمَثِّلُ عائقًا كبيرًا أمام عملية إحصاء المفردات الصينية داخل الذخائر اللغوية. ولذلك فعند عمل معالجة آلية للغة الصينية الطبيعية تكون البداية بعمل تمييز لحدود الكلمات داخل الذخائر الصينية. وفي تلك الفترة كانت جميع الدراسات العلمية التي تسعى إلى تصميم برمجيات للتمييز الآلي لحدود المفردات الصينية تستخدم منهجيات مختلفة، ولن يتم تجربة تلك البرمجيات بصورة عملية على الواقع اللغوي، ولا يمكن القول: إن هذه التجارب قد خرجت من إطار النظرية إلى حيز التطبيق، ولا يمكن الادعاء أن مشكلة التمييز الآلي للمفردات الصينية قد تم حلها بصورة كاملة، فما زال هناك نقاط عديدة في تلك البرمجيات تحتاج إلى تعديل وتحسين في الأداء. إن مشروع إنشاء هذه الذخيرة اللغوية الذي بدأ عام ١٩٧٩ قد تم في

ظل تلك الظروف. وفي ذلك الوقت تم الاستعانة بالعنصر البشري في إتمام عملية تمييز حدود المفردات الصينية، بالإضافة إلى وضع رموز خاصة تُحَدِّد خصائص كل عينة لغوية، وفي النهاية تم الاستعانة بإمكانات الحاسب الآلي في إنهاء عمليات التوثيق، والإحصاء، والتحليل، والتجميع وما إلى ذلك من أعمال.

أولاً: المهام الإحصائية على مستوى المفردات في اللغة الصينية الحديثة تشمل ما يلي:

أ- تصنيف المفردات وتجميعها، ورصد عدد مرات تواتر المفردات، والعدد الإجمالي لأنواعها، والعدد التراكمي لها، وحساب معدل التكرار النسبي والتراكمي للمفردات.

ب- حساب معامل انتشار كل نوع من أنواع المفردات الصينية داخل الذخيرة، ومؤشرات الاستخدام.

ج- تحديد درجات الاستخدام لكل المفردات وفقًا لمعدلات تكرارها داخل النصوص الصينية، وحساب متوسط طول المفردات الصينية مع تحديد عدد أنواع الكلمات المتواجدة في كل درجة وطرق توزيعها داخل كل طول من أطوال الكلمات.

د- حساب أعلى قيمة مشاهدة وأعلى قيمة مُحتَمَلة داخل الكلمات ذات معدل التكرار الأعلى.

هـ- إظهار نسب تواجد المفردات أحادية المقطع الصوتي، وثنائية وثلاثية ورباعية المقطع وما هو أكثر من ذلك داخل النصوص الصينية.

و- وضع قائمة بترتيب المفردات حسب درجة تكرارها.

ز- وضع قائمة بترتيب المفردات حسب درجات الاستخدام.

ح- وضع قائمة للمفردات ذات معدل التكرار الأقل؛ مثل المفردات التي تكررت أقل من خمس مرات، وأقل من عشر مرات.

ثانيًا: المهام الإحصائية على مستوى الرموز الصينية تشمل ما يلي:

أ- إظهار العدد الإجمالي لأنماط الرموز الصينية، وتوليد فهرس إجمالي للرموز الصينية.

ب- إحصاء عدد مرات تكرار الرموز الصينية في كل نمط من الأنماط، وحساب معدل التكرار النسبي ومعدل التكرار التراكمي لكل نمط.

ج- إحصاء عدد مرات اشتراك كل نمط من الرموز الصينية في تكوين المفردات، والمواقع المختلفة التي يحتلها داخل الكلمة (بداية الكلمة، ووسط الكلمة، وذيل الكلمة)

د- وضع قائمة للرموز الصينية حسب معدلات تكرارها، ووضع قائمة للرموز الصينية حسب عدد مرات تواترها داخل المفردات الصينية.

٢- الذخيرة اللغوية التي أنشأتها جامعة الطيران الصينية بالتعاون مع عدد من الجامعات

في التاسع عشر من شهر نوفمبر عام ١٩٨١، كَلَّفَت أكاديمية البحث العلمي الصينية هيئة التوحيد القياسي بمَهَمَّة عمل "إحصاء عدد مرات تكرار المفردات في اللغة الصينية الحديثة"، وقد تولت جامعة الطيران الصينية رئاسة هذا المشروع بمساعدة كل من جامعة الشعب الصينية، وجامعة بكين، وجامعة وو خان وغيرها من عشر جامعات صينية أخرى. وقد تم الانتهاء من هذه المهمة عام ١٩٨٦، وفي الثلاثين من يونيه عام ١٩٨٦ حصل مشروع "إحصاء عدد مرات تكرار المفردات في اللغة الصينية الحديثة" على تمويل من الحكومة الصينية. وفيما يلي تعريف بالأعمال التي تم إنجازها من خلال المشروع [42].

أ) مكونات الذخيرة

تنحصر حدود المادة اللغوية التي تم اختيارها لمشروع إحصاء معدل تكرار المفردات في اللغة الصينية الحديثة في المطبوعات الرسمية في الفترة ما بين ١٩١٩-١٩٨٢، وقد تم تقسيم تلك الفترة إلى أربع فترات زمنية فرعية، الفترة الأولى (١٩١٩-١٩٤٩)، والفترة الثانية (١٩٥٠-١٩٦٥)، والفترة الثالثة (١٩٦٦-١٩٧٦)، والفترة الرابعة (١٩٧٧-١٩٨٢). وفي كل فترة من الفترات سالفة الذكر، تم جمع المادة اللغوية من خلال فرعين لغويين أساسيين هما فرع العلوم الاجتماعية وفرع العلوم الطبيعية، وكل فرع من تلك الأفرع الرئيسة تم تقسيمه إلى خمسة تخصصات فرعية. وللتعرف على مزيد من المعلومات حول مُسَمَّيَات تلك التخصصات انظر الجدول رقم ٢-٩.

جدول (٢-٩): توزيع المادة اللغوية بمشروع الذخيرة اللغوية لجامعة الطيران ببكين

التصنيف		المحتوى	عدد الكلمات
العلوم الاجتماعية	١	الحياة المدنية (الأزياء، والطهو، والسياحة، وجمع طوابع البريد.. إلخ)	١١٤٧٤٨٢
	٢	التاريخ والفلسفة (علم النفس، وعلم التربية، وعلم الجمال، وعلم الاجتماع.. إلخ)	٢٥٥٦٨٠٤
	٣	السياسة والاقتصاد (التجارة، والمال، والإحصاء، والإدارة.. إلخ)	٢٧٤٧٩٦٠
	٤	الأخبار والتقارير (التقارير العسكرية الصادرة عن جيش التحرير الصيني)	٢٨٨٠٦١٥
	٥	الآداب والفنون (القصة، والنثر، والشعر، والمسرح، والأغاني)	٤٨٩٥٧٢٤
العلوم الطبيعية	١	العمارة والنقل (بما في ذلك البريد)	٥٩٧٠٤٦
	٢	الزراعة، والغابات، والرعي، وصيد الأسماك	١٢١١١٥٢
	٣	الصناعات الخفيفة (الإلكترونيات،و الصناعات الكيماوية لمستلزمات الحياة اليومية، والبلاستيك، والأطعمة، والنسيج.. إلخ)	١٤٤٨٤٤٥
	٤	الصناعات الثقيلة (المناجم، والتعدين، وصناعة الماكينات، والطاقة.. إلخ)	١١٧٧٥٨٦
	٥	المعارف والعلوم الأساسية (الرياضيات، والفيزياء، والكيمياء، والأحياء، والفلك، والجغرافيا، إلخ)	٢٤٢١٨٠٢

وقد تم الحصول على المواد اللغوية من المصادر التالية:

أ- الجرائد والمجلات.

ب- المقررات الدراسية.

ج- المؤلفات.

د- المطالعات العامة (بما في ذلك المؤلفات في العلوم الأساسية).

ولا تتضمن المواد اللغوية السابقة الأعمال المُتَرْجَمة؛ حيث إن هذه الأعمال تُركِّز على إنتاج الكُتَّاب المشهورين، بالإضافة إلى المؤلفات النموذجية في اللغات الأخرى، كما أن عدد المؤلفات في العلوم الطبيعية كان منعدمًا في خلال الفترة الأولى (١٩١٩-١٩٤٩)؛ وذلك لعدم إمكانية العثور على مؤلفات في العلوم الطبيعية في تلك الفترة باللغة الصينية.

وقد وصل حجم المادة اللغوية التي تم جمعها لمشروع معدل تكرار المفردات في اللغة الصينية الحديثة ثلاثمائة مليون رمز صيني، وقد تم اختيار المادة اللغوية من خلال عدة أساليب لجمع العينات، كالعينات العشوائية والعينات المنتظمة (مثل عينات الفترة- والعينات الطبقية) وقد وصل الحجم الإجمالي للعينات المختارة إلى ما يقرب من ٢٥ مليون رمز صيني.

ب) استخدامات الذخيرة

بعد بناء الذخيرة، قام فريق العمل بتنفيذ الأعمال التالية:

١- تقسيم المواد اللغوية في الفترة ما بين ١٩١٩ و١٩٨٢ إلى أربع فترات، كل فترة تضم مواد تمثل العلوم الاجتماعية، وأخرى للعلوم

الطبيعية، وكل فرع منها ينقسم إلى خمسة تخصصات فرعية، وقد تم عمل إحصاء للمفردات على كل قسم من تلك الأقسام بمفردها، وفي الوقت ذاته تم عمل إحصاء بمعدل تكرار استخدام علامات الترقيم الصينية التي تصل إلى أربع عشرة علامة.

٢- استخدام برمجية التمييز الآلي لحدود المفردات الصينية المعروفة باسم **CDWS** للمرة الأولى.

٣- تصميم برمجية حاسوبية متكاملة وإطلاقها لعمل الإحصاء اللغوي للكلمات في اللغة الصينية الحديثة.

٤- تأليف معجم إلكتروني يتكون من ١٣١١٦١ مادة لغوية.

٥- بناء بنك معلومات عن الرموز الصينية يحتوي ٥٢ خاصية من خصائص الرموز الصينية.

٦- طباعة إحصاء عام عن معدلات تكرار المفردات في الفترات الزمنية الأربع ١٩١٩-١٩٤٩، و١٩٥٠-١٩٦٥، و١٩٦٦-١٩٧٦، و١٩٧٧-١٩٨٢، بالإضافة إلى إحصاء عام عن معدلات تكرار المفردات الصينية على مستوى العلوم الاجتماعية والعلوم الطبيعية. وقد تم طباعة المخرجات السبع للمشروع في نسختين، الأولى حسب الترتيب الأبجدي لنطق الرموز الصينية، والثانية حسب معدل التكرار داخل الذخيرة. وقد وصل حجم المخرجات الورقية إلى عشرة آلاف صفحة.

وتَمَيَّز هذا المشروع في ذلك الوقت بالخصائص التالية:

أ- تم النظر إليه باعتباره المشروع الإحصائي ذي النطاق الأوسع. في ذلك الوقت لإحصاء معدل تكرار المفردات الصينية؛ وذلك على مستوى الحجم وسعة انتشار المادة اللغوية على مدى زمني طويل وشمولية التخصصات التي جُمِعَت على أساسها المادة اللغوية.

ب- تجانُس التوزيع الكمي للعينات، وضآلة التأثر بالخلفية الاجتماعية للنصوص، بالإضافة إلى دقة النتائج الإحصائية.

ج- لأول مرة يتم عمل تمييز آلي لحدود المفردات في اللغة الصينية الحديثة على مستوى الصين.

د- لأول مرة يتم استخدام الأكواد المختلطة لضغط الرموز والكلمات الصينية الأمر الذي أتاح إمكانية تمييز الرموز الصينية متعددة النطق، الأمر الذي جعل نتيجة الإحصاء أكثر دقة.

تم إنشاء الذخيرتين السابقتين على وجه الخصوص لعمل إحصاء كمي للمفردات والرموز الصينية، ومن خلالها تم استقراء المعلومات الخاصة باستخدام المفردات والرموز الصينية. إلا أنه للأسف لم تستمر عملية الاستفادة منهما في أعمال أخرى؛ بسبب عدم توحيد معايير التكويد الخاصة بالرموز فيهما، الأمر الذي جعل الاستفادة من هاتين الذخيرتين تتوقف عند حد مَهَمَّة الإحصاء اللغوي، وعدم استمرارية العمل بهما فيما بعد. وبذلك نكون قد عرضنا في النقطتين السابقتين أشهر ذخيرتين لغويتين للاستخدام العام في اللغة الصينية.

ثانيًا: الذخيرة المتوازنة لمعهد الدراسات المركزي بتايوان

الذخيرة المتوازنة لمعهد الدراسات المركزي بتايوان (يطلق عليها اختصارًا ذخيرة معهد دراسات تايوان المعروفة باسم Sinica Corpus) وهي أول ذخيرة صينية متوازنة تحتوي ترميزًا للأنواع النحوية للمفردات الصينية على مستوى العالم. وكان الهدف العام من تلك الذخيرة هو بناء ذخيرة متوازنة للغة الصينية تتكون من خمسة ملايين رمز صيني.

١- الفكرة من وراء بناء ذخيرة معهد الدراسات المركزي بتايوان

سعى فريق العمل المُكلَّف ببناء ذخيرة معهد الدراسات المركزي بتايوان إلى جمع المادة اللغوية للذخيرة منذ عام ١٩٩٠ (Huang & Chen:1992)[44]، وقد تم جمع نصوص صينية حديثة فيما يقرب من عشرة ملايين رمز، ونصوص صينية كلاسيكية تزيد عن مليون رمز (Huang & Chen:1994)[45]. ونظرًا إلى الكفاءة التي يتمتع بها الفريق في معالجة المواد اللغوية الصينية، بالإضافة إلى خبرتهم في معالجة الذخائر اللغوية الإلكترونية كبيرة الحجم (K.J.Chen:1996)؛ فقد تمتع الفريق بعناصر قوة حقيقية، وضم كفاءات بشرية مكنته من الاضطلاع بمَهَمَّة بناء ذخيرة متوازنة للغة الصينية. وقد استهدف فريق البحث في البداية بناء ذخيرة لغوية حجمها مليوني رمز صيني، ثم سعى إلى تحقيق المستهدف النهائي وهو الوصول بحجم الذخيرة إلى خمسة ملايين رمز بعد عدة سنوات، ومن ثَمَّ الاقتراب من نطاق الذخائر اللغوية العامة التي تُبْنى في حقل اللغويات الحاسوبية. وتظهر الفكرة من وراء بناء هذه الذخيرة اللغوية من خلال النقاط الثلاث التالية:

أ) الالتزام بالمعايير القياسية لجمعية اللغويات الحاسوبية بتايوان في تمييز حدود الكلمات الصينية .

إن تمييز حدود الكلمات الصينية من الشروط الأساسية التي تبدأ بها عملية معالجة اللغات الطبيعية باللغة الصينية. ولكن نظرًا إلى عدم وجود مسافات فاصلة بين حدودى المفردات بعضها البعض في المؤلفات التحريرية الصينية، بالإضافة إلى وجود خلافات حول تلك الحدود الفاصلة والمحَدَّدَة للمفردات في اللغة الصينية من حيث وضع تعريف واضح وموحد لحدود الكلمة في اللغة الصينية، فقد تم الاتفاق بين أعضاء فريق بناء الذخيرة على اتباع مقاييس مَجْمَع اللغة الصينية بتايوان في تحديد قواعد تمييز حدود الكلمات الصينية. وهذا القرار ليس من شأنه فقط تشجيع المشاركة العامة في المادة اللغوية المُكَوِّنَة للذخيرة، بل من شأنه أيضًا تسهيل الحصول على تغذية راجعة مستمرة من مستخدمي الذخيرة عن آرائهم في الطريقة المُتَّبَعة في تقسيم المفردات داخل الذخيرة، الأمر الذي ينتج عنه مرجعية مُهِمة لتصحيح مسار تمييز حدود الكلمات الصينية داخل النصوص فيما بعد.

ب) عند اختيار العينات للذخيرة تم اعتماد وحدة الإختيار لتكون اكتمال الفقرات بشكل طبيعي، وليس طول النصوص.

فقد كان أحد المعايير المُتَّبَعة داخل ذخيرة براون هو السعي إلى تحقيق التوازن في توزيع المادة اللغوية على التخصصات المختلفة، وقد أدى ذلك إلى اعتماد طول موحَّد لاختيار العينات العشوائية في الذخيرة؛ بحيث يكون طول كل نص ٢٠٠٠ كلمة. وقد رأى فريق العمل في ذخيرة معهد

الدراسات المركزي بتايوان أن اتباع تلك الطريقة من شأنه أن يتسبب في عدم اكتمال المضمون اللغوي للعينات. وبالإضافة إلى ذلك، فإن طول النص دائمًا ما يعبر عن إحدى الخصائص المهمة في النص، فإذا تم اختيار النصوص على اعتبار طول مُوَحَّد من حيث عدد الكلمات، فإن ذلك يُفْقِدُها هذه الخاصية. ولذلك، فعلى الرغم من أن ذخيرة معهد الدراسات المركزي بتايوان قد تجنبت اختيار النصوص الطويلة جدًّا والنصوص القصيرة جدًّا، فإنها لجأت بعد تحديد النص إلى معيار أخذ العينات وفقًا للحدود الطبيعية للفقرات. فقد كانوا يعتقدون أن هذا التصرف يُمَكِّنُهم من الحصول على نصوص ذات معلومات لغوية متكاملة.

ج) استخدام عدة مستويات لتمييز حدود الكلمات في المادة اللغوية

نظرًا إلى تعدد العوامل التي تؤثر في ملامح اللغة بصفة عامة، فإن النظر إلى اللغة من خلال خاصية معينة مثل الموضوع وأسلوب الكتابة في تحديد الذخيرة المتوازنة يُعْتَبَر أمرًا غير كاف. ورغبةً في التغلب على النظرة السطحية في توصيف المادة اللغوية، فقد لجأ فريق بناء الذخيرة إلى خمسة معايير مختلفة يتم النظر إلى العينات من خلالها هي: أسلوب الكتابة، والفكرة، والبناء اللغوي، والعنوان، ووسيط النشر. وعلى الرغم من أن فريق العمل قد لجأ إلى أن يكون العنوان هو المعيار الأساس في تحقيق التوازن في الذخيرة، فإنهم كانوا يأملون بعد تحقيق عدد من النتائج، في إمكانية استخدام أكثر من معيار لتحسين مستوى التوازن في ذخيرتهم، على أن يكون ذلك في مرحلة لاحقة من العمل. وهناك فائدة أخرى من هذا الأسلوب، ألا وهي

تسهيل العمل في بناء الذخيرة. وذلك على أساس أن الباحثين الذين سيستخدمون الذخيرة فيما بعد بإمكانهم اختيار مجموعة من المعايير يتم العمل على أساسها في توليد ذخائر فرعية خاصة بهم، كما يمكنهم عمل دراسات مقارنة على تلك الذخائر الفرعية.

٢- تصنيف المادة اللغوية وطرق اختيار النصوص في الذخيرة المتوازنة

من أجل التوصل إلى إدارة متميزة في عملية اختيار المادة اللغوية بصورة متوازنة، أُرفِقَت مع كل وحدة لغوية علامات خاصة توضح أسلوب الكتابة، والفكرة، والبناء اللغوي، والعنوان، ووسيط النشر، بالإضافة إلى اسم المؤلف، ونوعه، وجنسيته، ودار النشر، وما إلى ذلك من خصائص.

أ) تحديد خصائص النصوص

بعد الاطلاع على الخبرات السابقة في إدارة كل من ذخيرة لوب، وبراون، وكوبويلد، فقد تم الاستعانة بالمبادئ المُتَبَعَة في تصنيف الكتب، وقد حدد فريق العمل قائمة من الخصائص على أساسها صُنِفَت الكتب الصينية. وهذه الخصائص يتم الاستعانة بها في توضيح مصدر المادة اللغوية، وأسلوب الكتابة، والمضمون الذي تدور حوله المادة اللغوية وغير ذلك، فالعنوان يُعَبِّر عن مضمون النص، وتصنيفه، أما أسلوب الكتابة فيوضح طريقة التعبير عن النص. بالإضافة إلى وضع علامات خاصة توضح ثلاثة مستويات معلوماتية هي وسيط النشر، والمؤلف، ودار النشر. فوسيط النشر، يُوَضَّح من خلال تحديد مصدر المادة اللغوية. والمعلومات الخاصة بالمؤلف توضَّح من خلال كتابة اسم المؤلف، ونوعه، وجنسيته، ولغته الأم. ودار النشر توضَّح من

خلال تسجيل المعلومات الخاصة بالطباعة كاسم دار النشر، وتاريخ النشر، وعدد الطبعات.

ب) الموضوع

يتحدد الموضوع من مضمون النص وصُلْب القضية التي يناقشها. وقد استخدم فريق إنشاء الذخيرة منهجية تصنيف الكتب في علم المكتبات كأساس لتحديد خصائص الموضوعات.

ج) أنواع النصوص

أنواع النصوص يُقْصَد بها أسلوب كتابة النص؛ حيث تنقسم النصوص إلى تقارير، ومناقشات، وإعلانات، ومراسلات، ومكاتبات رسمية، وقصص، وأساطير، ونثر، ومذكرات، وسيرة ذاتية، وشعر، وسجلات، وكتالوجات، ومسرحيات، وحوارات، وخطب، وسجلات المؤتمرات. وقد استعان المصممون بالنصوص الاستشهادية التي ترد على هوامش المجلات والجرائد المستخدمة ولكن بنسبة ضئيلة جدًّا. أما المراسلات فقد كانت من ثلاثة مصادر مختلفة: مراسلات القراء للجرائد والمجلات، ونماذج المراسلات المرفقة في المقررات الدراسية العلمية، والمراسلات الإلكترونية. أما النصوص المسرحية فمصدرها جميعًا من دروس مرحلة التعليم الابتدائية، وجميعها من النصوص النثرية وموضوعاتها في إطار أدب الأطفال، والشكل اللغوي لها عبارة عن لغة شفهية مدونة. أما الخطب فتضم مسودات الخطب حول المبادئ الثلاث للديمقراطية، بالإضافة إلى بعض الخطب التي تحولت إلى كتب أو المنشورة في المجلات.

د) وسيط النشر

وسيط النشر يتم تقسيمه حسب مصدر المادة اللغوية إلى وسيطين مختلفين هما: الوسيط التحريري والوسيط الشفهي. فالوسيط التحريري للمادة اللغوية يمكن تقسيمه إلى: دوريات، وكتب، ومراسلات، ووسائط مرئية وسمعية، ومؤتمرات، ومتفرقات؛ وتضم الوسائط المرئية والسمعية الحوارات التي تستخدم في البرامج التليفزيونية بين سيدتين، بالإضافة إلى النصوص المستخدمة في النسخة الإلكترونية من المستندات الحكومية. وقد ساعد هذا النوع من النصوص بصورة كبيرة في جمع عدد هائل من المادة اللغوية للذخيرة؛ حيث لم تكن هناك حاجة لإضاعة وقت كبير في الحصول على حقوق النشر، بالإضافة إلى عدم الحاجة إلى تدقيق تلك النصوص من حيث التحويل الخاطئ للكلمات، وغير ذلك من المشكلات التي توجد في الأشكال الأخرى من النصوص، بالإضافة إلى إمكانية جمع أنواع عديدة من النصوص بما يحقق تنوع المادة اللغوية. وإذا كانت المستندات الحكومية الإلكترونية بها إشارة إلى مصدر المستند، فيمكن تصنيفها مباشرة حسب نوع وسيط النشر الخاص بها. أما النصوص التي لا تحتوي تلك المعلومات فلا تُدرَج تحت أي تصنيف. أما الدوريات فتنقسم إلى جرائد، ومجلات علمية، ومجلات عامة، وتنقسم الكتب إلى كتب تعليمية، وكتب مساعدة، ومؤلفات علمية، ومؤلفات عامة. أما الدوريات فتضم الجرائد الدورية الصينية وغير الدورية، بالإضافة إلى الجرائد اليومية للأطفال، ونشرات مركز الحاسب بمعهد الدراسات المركزي بتايوان. وتضم المجلات العامة مجلات الشئون

العامة والموضة، والسياحة، ومجلات الشعر، ومجلات السينما؛ أما المجلات العلمية فتضم معلومات مختصرة عن الأطباء، والمجلات الصادرة عن وكالة القوميات. أما الكتب التعليمية فتضم كتب الأدب واللغة في المرحلة الابتدائية، والحوارات المستخدمة باللغة الصينية الصادرة عن مركز المعلمين بتايوان؛ أما الكتب المساعدة فقد جَمَعت في إطارها التقارير الفنية الصادرة عن فريق بناء ذخيرة المفردات الصينية. أما المؤلفات العلمية فتشمل الأبحاث العلمية. وتشمل الكتب العامة مسودات الخطب السياسية عن المبادئ الثلاثة للديمقراطية، والمؤلفات الشهيرة في علم النفس الذي تم تمويلها من صندوق خونغ تيان تشوان وما إلى ذلك. والمادة اللغوية الشفهية مصدرها الحوارات اليومية لطلاب الصين الأم المغتربين في أميركا.

هـ) أسلوب الكتابة

تنقسم النصوص من حيث أسلوب الكتابة إلى الأسلوب السردي، والأسلوب الجدلي، والأسلوب التفسيري، والأسلوب الوصفي. فالأسلوب السردي هو الذي يتناول وصف الإنسان والأشياء من حيث الخصائص والحركة والتغيرات التي تحدث، فالمقالات التي تتناول سردًا عامًا أو تلك التي تقدم تقاريرًا معلوماتية تم إدراجها تحت تصنيف الأسلوب السردي. وقد كانت النصوص السردية هي أكثر الأساليب المستخدمة من بين ما تم جمعه من نصوص الذخيرة. أما النصوص الجدلية فهي تلك التي تَعْرِض رأي الكاتب الشخصي وتسعى إلى الحصول على تأييد الآخرين وإقناعهم بوجهة نظر معينة. والأسلوب العلمي وظيفته الأساسية هي تحليل الظواهر المترتبة

على حدث معين، أو إظهار الحقائق العلمية، الأمــر الــذي يمــد المتلقــي بمعلومات أو حقائق عن شيء معين. ولذلك فإن النــصوص التــي تتنــاول وظائف وطبيعة الأشياء من وجهة نظر موضوعية تنتمــي إلــى الأســلوب العلمي. أما الأسلوب الوصفي فذلك الذي يتناول الأشــخاص، والحيوانــات، والأشياء أو الأماكن بالوصف المتعمق، ومن الممكن أن يحتوي ذلك الأسلوب تشبيهات واستعارات وكنايات وغير ذلك من الأساليب البلاغية؛ حيث يستعين بها الكُتّاب في الوصف العميق لموضوع ما وإظهــار خصائــصه بــصورة دقيقة. النصوص الوصفية تحتوي أيضًا ما يتعلق بالمشاعر الداخلية للكاتــب مثل اليوميات التي تصف المشاهد والكائنات.

و) الشكل اللغوي

الشكل اللغوي يعني الأسلوب الذي تظهر عليه لغــة النــصوص فــي الذخيرة، حيث تنقسم مادة الذخيرة إلى لغة تحريرية ولغــة شــفهية، ومــن المعروف أن هذين الشكلين يختلفان بصورة كبيرة. ويمكننا تقسيم اللغة مــن حيث الشكل إلى لغة مكتوبة (**written**)، ولغة مكتوبة لكي تُقْرَأ (**written-to-be-read**)، ولغة مكتوبة لكي يتم التحــدث بهــا (**written-to-be spoken**)، ولغة منطوقة **spoken**، ولغة منطوقة لكي تُكْتَب (**spoken-to-be-written**). أما اللغة المكتوبة فتشير إلى اللغة التحريرية العامة، ويندرج تحتها غالبيــة النصوص التي تم جمعها في الذخيرة، أما اللغة المكتوبة لكي يتم التحدث بها فتشمل النصوص المسرحية والحوارات الإذاعية التي تُكْتَب لكي يــستخدمها شخص ما في ظروف معينة؛ ولذلك فهي تختلف إلى حدٍّ كبيــر مــع اللغــة الشفهية التي تُستَخدم في الحياة الواقعية؛ أما اللغة المنطوقــة فتــشير إلــى

الحوارات الشفهية العامة، ويعتبر جمع هذا النوع من المادة اللغوية وتنظيمها صعب إلى درجة كبيرة، لذلك فإن نسبة تمثيله داخل الذخيرة جاءت قليلة.

أما اللغة المنطوقة لكي تُكتَب فتشير إلى ذلك النوع من النصوص الذي يتم تدوينه داخل المؤتمرات؛ ونظرًا إلى وجود فرصة للترتيب والمعالجة، فيمكن حذف العديد من العناصر اللغوية الزائدة التي تمثل طبيعة اللغة الشفهية داخل وقائع المؤتمر؛ وذلك من أجل أن يتم التمييز بينها وبين كل من اللغة الشفهية أو اللغة التحريرية الواقعية.

٣- نسب النصوص المكوِّنة لذخيرة معهد الدراسات المركزي بتايوان

تم توزيع المادة في ذخيرة معهد الدراسات المركزي بتايوان بصورة رئيسة حسب الموضوعات، وقد وصلت نسبة تمثيل تلك الموضوعات داخل الذخيرة حتى الآن إلى ما يلي:

الفلسفة: ١٠%

العلوم: ١٠%

الاجتماع: ٣٥%

الفنون: ٥%

الحياة: ٢٠%

الأدب: ٢٠%

وقد تم اختيار المادة اللغوية وفقًا لتلك النسب من المصادر التالية مع اعتبار أن وحدة العد هي الرموز الصينية:

أ) الصُحُف

الصحف الصينية: ٥٠٠٥٦٦ رمزًا، الصحف الحرة: ١٢٥٨٣٣٤ رمزًا، صحف الأطفال: ٢٩٩٢٦٠ رمزًا، نشرات ومراسلات مركز الحاسب التابع لمعهد الدراسات المركزي بتايوان: ٩٥٧٧٤٢ رمزًا.

ب) المجلات العامة

مجلات الشئون العامة: ٦١٩٩٤ رمزًا، مجلات الموضة: ٢٩٨٤٠ رمزًا، النشرات السياحية: ١٣٨٤٦٢ رمزًا، مجلات السينما العالمية ١٤٨٦٩ رمزًا.

ج) الدوريات العلمية

الدوريات الصادرة عن مركز القوميات التابع لمعهد الدراسات المركزي: ١١٢٢٥ رمزًا، نشرات الأطباء بمعهد الدراسات المركزي: ٣٩٥٠٧ رمزًا.

د) الكتب التعليمية

المؤلفات الدراسية الصادرة عن الحكومة لمرحلة التعليم الابتدائي، التي تقع في ١٢ مجلدًا: ٨٨٧٤٤ رمزًا.

هـ) الكتب المساعدة

التقرير الفني لفريق بناء ذخيرة المفردات التابع لمركز المعلومات بمعهد الدراسات المركزي: ٣٨٤٢ رمزًا.

و) المؤلفات العلمية، والأبحاث

٣٩٠٧٦ رمزًا.

ز) النصوص التي لا يمكن إدراجها تحت تصنيفات سابقة

٨٠٠١ رمز.

ح) الكتب

عدد ثمانية مجلدات لعلم النفس صادرة عن جمعية صندوق خونغ تيان تشوان.

ط) الوسائط المرئية والمسموعة

المقالات الصادرة عن الدوائر العلمية التايوانية: ١٠٣٩٥٥ رمزًا.

ي) حوارات ومقالات صحفية:

٢٨٨٣١ رمزًا.

٤- تمييز حدود الكلمات داخل الذخيرة وترميزها من حيث التصنيف النحوي

اتبعت ذخيرة معهد الدراسات المركزي بتايوان أسلوب التمييز الآلي لكلمات الذخيرة. وقد كانت عملية التقسيم تتم اعتمادًا على التسعين ألف مادة لغوية التي تمثل محتويات المعجم الصادر عن معهد الدراسات المركزي بتايوان. أما المفردات التي لم ترد في المعجم، فكان التعامل معها يتم باعتبارها رموزًا ويتم تمييزها على أساس الرمز وليس الكلمة. وبعد ذلك يتم الجمع بين الرموز التي لها قوة توليد كلمات عن طريق اللواحق والرموز المعبرة عن الأرقام في إطار كلمات مركبة. وعند تمييز الكلمات، كانت المبادئ العامة للتمييز كما يلي:

أ) الوحدات اللغوية التي لها استقلال من حيث المعنى، يتم اعتبارها وحدة يتم تمييز الكلمات على أساسها.

ب) يتم تمييز الوحدات اللغوية الشائعة وفقًا لما هو متعارف عليه في الاستخدام اللغوي.

ج) المعنى والتركيب يتم النظر إليهما معًا باعتبارهما أساس تمييز حدود الكلمات.

د) يتم التمييز على أساس العلامات التي لها مغزى واضح في الفصل بين حدود الكلمات.

هـ) الوحدات اللغوية المتفِقَة من حيث الشكل والمختلفة من حيث التركيب يتم تمييزها على أساس السياق اللغوى.

و) إذا كان هناك تضارب في مبادئ التمييز يتم اللجوء إلى اتفاقية التوحيد المعياري الصادرة عن جمعية اللغويات الحاسوبية.

وبناءً على ما سبق تم تمييز حدود الكلمات في ذخيرة معهد الدراسات المركزي في تايوان، وقد تم اللجوء إلى منهجية البدء بالحاسب، ثم المراجعة من خلال العنصر البشري في عمل الترميز اللغوي لمحتوى الذخيرة بالكامل، وقد تم استخدام عدد ستة وأربعين رمزًا في عملية الترميز النحوي للمفردات.

وبالإضافة إلى عملية ترميز أقسام الكلام، فقد تم عمل ترميز آخر لبعض العناصر النحوية الخاصة بالذخيرة، وقد تم استخدام ثماني علامات

لهذا الغرض، وقد صُمِّمَت هذه العلامات خصيصى من أجل الاستخدام في رصد بعض الظواهر اللغوية في اللغة الصينية؛ مثل ظاهرة الفعل والمكمل، والأفعال المتعدية، والوحدات اللغوية التي تتخلل وحدات متعارف عليها سابقًا، والتراكيب الاسمية والكلمات الدخيلة.

علاوة على ذلك، فقد تم إضافة خاصية الاستعلام داخل الذخيرة وفقًا للكلمات المفتاحية؛ حيث يمكن استخراج السياقات التي تحتوي كلمات محددة من بين محتوى الذخيرة، بالإضافة إلى تقديم المعلومات الخاصة بتمييز حدود الكلمات وترميزها.

ويُظْهِر الجدولان ٢-١٠ و٢-١١ نتيجتين للاستعلام بكلمة مفتاحية[١] داخل ذخيرة معهد الدراسات المركزي بتايوان إحداهما تعرض مجموعة سياقات وردت فيها الكلمة المفتاحية دون إضافة معلومات عن التصنيف النحوي للكلمات، والنتيجة الثانية السياقات نفسها بعد تمييز حدود الكلمات، ومُرفقًا مع كل كلمة الترميز الخاص بتصنيفها نحويًّا[٢].

---

(١) الكلمة المفتاحية التي تم الاستعلام عن سياقاتها هي كلمة "試試" "يُجَرِّب"، وقد تمت كتابة هذه الكلمة بالبنط السميك ووُضِعَ أسفلها خط في النص المترجم لتسهيل عملية تمييزها على القارئ العربي.(المترجم)

(٢) الغرض التواصلي من عرض الجدولين ٢-١٠ و٢-١١ في النص الأصلي هو إظهار نتيجة الاستعلام بكلمة مفتاحية داخل الذخيرة؛ مرة عندما تكون النتيجة دون ترميز، ومرة ثانية عندما تكون النتيجة مُرمَّزة. ويلاحَظ أن النتيجة المرمزة لم تشمل جميع الكلمات في الاستعلام الأول، بل اقتصرت على الكلمات المُخزَّنة في برمجية الترميز. ونوجه عناية القارئ إلى أن الكتاب الأصلي ضمَّ مُرفقًا به مجموعة العلامات المستخدمة في ترميز اللغة الصينية ومعناها، وقد قمنا بترجمة ما تشير إليه تلك العلامات باللغة العربية.(المترجم)

جدول (٢-١٠): مواد لغوية من تسعة عشر مقطعـا من نصوص مختلفة

:" 既然你非和我比剑不可，我也乐于试试你的本事，不过，　我随身带着主君的
且又影响家庭生活甚巨　，　所以没有出去试试阔天下　，　老有遗憾。此时，不妨以
用过一次没什么感觉了下次再去买一副试试只是不便宜耶　明天还要考动力学我要去。
对非物理学者而言你可试试　Close　Sutton　和　Marten　的此书包含了许多
的绝世刀法。学了两年　，　懒残大师有意试试他的功力　，　便把他叫来禅室，其时外面
还在纽约的话　，　我们就去买双冰刀鞋来试试　，　你就说在屏东溜冰的故事，穿好那种
带动人际之间的热络气氛，你是否也想试试，以镖会友一番　？快加入飞镖行列吧！
自认年轻貌美　，　身材还过得去　，　不妨去试试运气。假设你很不幸地没有上述这些
相信的迷信疗法　，　她也抱着一线希望去试试　，　两年来她为了能传宗接代做个真正的
及钢琴等自动演奏乐器，游客也可试试身子，弹奏乐曲，　体验开演奏会的
了　，　你走吧　！　狐狸说:　象啊　！　老虎只是试试你的胆子大不大而已，没想到你的胆子，
-真是一举两得呀！各位有空时不防试试我的消暑妙方，不　然，你总不能一箭
为什么它的名字中有个蚤字　，　大家不妨试试身手　，

看能不能徒手抓到它们。图说:1
是不喝酒的，看看杯子这么可爱，也想试试。" 格林哥说:" 傻瓜，这不是杯子，
中国人上了床 ， 功夫特别好 ， 不信可以试试。 " 她们信以为真，极感趣地打量着
一句:"不论你同不同意，今夜我要试试。 "" 喂！中国人！ 难道你们连做爱都
那人关了门 ， 开始剥身上衣裳: 你看我试试夹袄合适不。直到那色鬼剥光衣裳露出
一些创意 ， 不是一举数得吗 ? 大家不妨试试。大台北经济证券，新台币汇率昨日再，
他刚学会开车 ， 回国后 ， 喜欢在台湾试试他的开车技术。没有开两条街，就大叫

جدول (٢-١١): بعض مفردات المواد اللغوية السابقة

مُرَمَّزة نحويًا

不可 (D), 我 (Nh) 也 (D) 乐于 (VL) 试试 (VF) 你 (M) 的 (De) 本事(Na), 不过 (Cbb)
巨 (VH), 所以 (Cbb) 没有 (D) 出去 (VA) 试试 (VF) 闯 (VC) 天下(Nc), 老 (D) 有 (V-2)
再 (D) 去 (D) 买 (VC) 一 (Neu) 副 (Nf) 试试 (VF) 只是 (D) 不 (D) 便宜(VH) 耶 (T)
学者 (Na) 而 (Cbb) 言 (VE) 你 (Nh) 可 (D) 试试 (VF)Close (FW) 和 (Caa) (FW)Sutton
年 (Nf), 懒残 (Nb) 大师 (Na) 有意 (VL) 试试 (VF) 他 (Nh)

的 (De) 功力 (Na), 便 (D) 把 (P)
买 (VC) 双 (Nf) 冰刀 (Na) 鞋 (Na) 来 (D) 试试 (VF), 你 (Nh) 就 (D) 说(VE) 在 (P) 屏东 (Nc)
你 (Nh) 是否 (D) 也 (D) 想 (VE) 试试 (VF), 以 (P) 镖 (Na) (Neu)一会 (VC) 友(Na)
还 (Dfa) 过得去 (VH), 不妨 (D) 去 (D) 试试 (VF) 运气 (Na)。假设(VE) 你 (Nh) 很 (Dfa)
着 (Di) 一 (Neu) 线 (Na) 希望 (Na) 去 (D) 试试 (VF), 两 (Neu) 年 (Nf)来 (Ng) 她 (Nh) 为 (P)
乐器 (Na), 游客 (Na) 也 (D) 可 (D) 试试 (VF) 身子 (Na), 弹奏 (VC)乐曲 (Na)
:象 (Na) 啊 (T)! 老虎 (Na) 只是 (D) 试试 (VF) 你 (Nh) 的 (De) 胆子(Na) 大 (VH) 不 (D)
位 (Nf) 有空 (VH) 时 (Ng) 不妨 (D) 试试 (VF) 我 (Nh) 的 (De) 消暑(VA) 妙方 (Na),
蚤 (Na) 字 (Na), 大家 (Nh) 不妨 (D) 试试 (VF) 身于 (Na), 看 (VE) 能不能 (D) 徒手 (D)
这么 (D) 可爱 (VH), 也 (D) 想 (VE) 试试 (VF)。" 格林哥 (Nb) 说(VE):"傻瓜 (Na)
好 (VH), 不 (D) 信 (VK) 可以 (D) 试试 (VF)。" 她们 (Nh) 信以为真(VH),
今夜 (Nd) 我 (Nh) 要 (D) 试试 (VF)。"" 喂 (I)! 中国人 (Na)! 难道(D)
衣裳 (Na): 你 (Nh) 看 (VE) 我 (Nh) 试试 (W) 夹袄 (Na)

合适 (VH)不 (T)。 直到 (P) 吗 (T)? 大家 (Nh) 不妨 (D) 试试 (VF) 。大台北 (Nc) 经济 (Na) 证券(Na) 后 (Ng), 喜欢 (VK) 在 (P) 台湾 (Nc) 试试 (VF) 他 (Nh) 的 (De) 开(VC) 车 (Na) 技术 (Na)

ثالثًا: الذخيرة التزامنية لخمس مناطق صينية

الذخيرة التزامنية لخمس مناطق صينية يُطلَق عليها اختصارًا ذخيرة LIVAC، والتسمية الكاملة لها هي: Linguistic Variety in Chinese Communities وقد تم إطلاق هذه الذخيرة من قِبَل جامعة المدينة بهونج كونج.

وقد اتبعت هذه الذخيرة أسلوب جمع المادة اللغوية من خلال خطوات تزامنية، والغرض من ذلك اختيار مادة لغوية مُمَثِّلَة لكل منطقة وبحجم مناسب، بغرض التوصل لوصف الاستخدام اللغوي الواقعي للغة الصينية في عدة مناطق على مستوى الصين، بالإضافة إلى تدعيم النظريات المتعلقة باللغة الصينية بالشرح.

ويقدم المرجع رقم [47] شرحًا تفصيليًا لهذه الذخيرة من حيث عمليات البحث والتطوير

١- خصائص ذخيرة ليفاك LIVAC

تظهر خصائص الذخيرة المذكورة من خلال النقطتين التاليتين:

أ) تعتبر عملا غير مسبوق على مدى التاريخ من حيث جمع مادة لغوية هائلة الحجم من خمس مناطق صينية بشكل تزامني. وتضم

هذه المناطق الخمس بالإضافة إلى الصين الأم كلا من هونج كونج، وتايوان، وسنغافورة، ومكاو.

ب) في أثناء تصميم الذخيرة كان هناك مثابرة وإصرار على جمع المادة اللغوية خلال فترة زمنية طويلة جدًا مع مراعاة مستوى التزامن في جمع النصوص على مستوى المناطق الخمس.

إن جوهر علم اللغة التزامني هو دراسة اللغة من خلال قضية افتراضية عبر فترة زمنية محددة؛ بغرض وصف حالة اللغة في تلك الفترة، مع عدم الالتفات إلى الوضع السابق والمستقبلي لتلك الحالة.

إن اللغة كائن حي، يتغير آلاف المرات مع مرور الزمن، وأهم مظاهر هذه الحالة من التغير هو ظهور المفردات الحديثة وانتشارها، أو تغير معنى المفردات الموجودة بالفعل أو اختفائها.

ولذلك فإن جمع المادة اللغوية تحت شروط معينة من الأفضل أن يكون تحت مظلة فترة زمنية معينة بدرجة أو أخرى، الأمر الذي يسمح بعمل مراقبة ومشاهدة لظروف انتشار وتغير عدد من المفردات، بما في ذلك احتمال خضوع تلك المفردات لتأثير البيئات المختلفة. ولذلك فإن ذخيرة ليفاك قد حددت المرحلة الأولى لجمع المادة اللغوية بثلاث سنوات.

وحتى يمكن الحصول على مواد لغوية مناسبة، تم توزيع المادة اللغوية الصحفية التي جُمِعت للذخيرة على المناطق الخمس؛ بحيث يتم الحصول عليها من صحف تلك المناطق من منشورات يوم واحد مع مراعاة أن تكون الموضوعات متشابهة قدر الإمكان.

والمقصود بالتزامن في ذخيرة ليفاك هو نوع من التزامن أكثر صرامة من ذلك المتعارف عليه في مفهوم التزامن المستخدم في علم اللغة التزامني؛ بمعنى أن يتم عمل مراقبة زمنية لمحتويات الذخيرة في خلال الفترة الزمنية التي يتم التعامل معها، والغرض من ذلك هو الرغبة في أن تكون هناك فرصة لمراقبة ظواهر النمو أو الاضمحلال لبعض المفردات بصورة كاملة، بالإضافة إلى التعرف على مصادر تلك المفردات ومغزاها.

٢- نطاق ذخيرة ليفاك وطرق جمع المواد اللغوية

بدأت الإرهاصات الأولى لفكرة ذخيرة ليفاك منذ عام ١٩٩١، ومع حلول عام ١٩٩٣ بدأ التخطيط الفعلي لها بعد الحصول على التمويل اللازم لبنائها. وتَمَثَّلت الخطوات التنفيذية لذلك في جمع المادة اللغوية عن طريق اختيار مواد صحفية صادرة في يوم واحد في المناطق اللغوية الخمس: هونج كونج، ومكاو، وشنغهاي، وسنغافورة، وتايوان، على أن يتم جمع المادة مرة كل أربعة أيام. تضم المادة المقالات الافتتاحية، وجميع الأخبار والمقالات في الطبعة الأولى، والكتابات والمناقشات الخاصة بالشئون الدولية والمحلية، على أن يكون مُجْمَل ما يتم جمعه كل يوم في حدود عشرين ألف رمز صيني. وفي العامين من شهر يوليو ١٩٩٥ حتى يونيو ١٩٩٧ وصل حجم المادة التي جُمِعَت لذخيرة ليفاك على التوالي ٦٧٨٠١ رمز، و٨٤٤٥٣ رمزًا.

٣- التمييز الآلي للكلمات وبناء الذخيرة اللغوية

الهدف الأساس من بناء ذخيرة ليفاك هو عمل تحليل للمفردات؛ لـــذلك فقد كانت المَهَمَّة الأولى في عملية بناء الذخيرة هي عمل تمييز آلي لحـــدود الكلمات داخل الذخيرة، وقد استَخدمت ليفاك منهجية أقصى طـــول لتَجَمُّـــع الرموز في تمييز حدود الكلمات الصينية؛ حيث يقوم الحاسب بعمل التقـــسيم الآلي للمفردات في البداية على أساسها، بالإضافة إلى إضافة بعض العلامات بصورة آلية إلى عدد من الكلمات الخاصة (مثل الأرقام، وأسماء الأشخاص، وأسماء الأماكن) وذلك لتسهيل عملية الاستخدام والتحليل فيمـــا بعـــد. وقـــد وصلت دقة التمييز إلى ما يزيد عن نسبة ٩٥%. وبعد ذلك يتم عمل مراجعة باستخدام العنصر البشري لنتيجة التمييز، ويكون ذلك من خـــلال خطـــوتين رئيستين: أولاً: عمل فحص لجميع النصوص التي تم عمل تمييـــز لكلماتهـــا وتصويبها، ثم استخراج قائمة تضم جميع المفردات المكوِّنة للذخيرة بصورة آلية؛ ثانيًا: فحص القائمة ومراجعتها لاكتشاف ما إذا كانت هناك مفردات قد تم تمييزها بصورة غير مناسبة، وفي النهاية يتم إدراج هذه المفـــردات فـــي معجم ليفاك؛ وذلك من أجل رفع مستوى دقة البرنامج الـــذي يقـــوم بتمييـــز المفردات في ذخيرة ليفاك.

أما المقالات التي يقوم بمراجعتها العنصر البشري، فيتم عمل قـــراءة آلية لها بعد ذلك باستخدام الحاسب الآلي، وتستخدم المعلومـــات التـــي يـــتم الحصول عليها في بناء ذخيرة ليفاك. وبالإضافة إلى تسجيل جميع المفردات، يتم تسجيل المنطقة التي وردت منها، وتاريخ النشر، بالإضافة إلى المعلومات

الأخرى التي لها فائدة؛ مثل النطق الصيني القياسي، والنطق المحلـي لهـا. ومن أجل تسهيل عملية الاستعلام داخل المواد التي تم جمعها، تـم تـدوين جميع محتويات النصوص وعناوينها داخل الذخيرة اللغوية. وبالإضافة إلـى ذلك، فقد تم تسجيل المكان، والتاريخ، والفقرة، والجملة وموقع ظهور الكلمة داخل الجملة وما إلى ذلك؛ وذلك من أجل زيادة سرعة البحث والتنقيب عـن البيانات داخل الذخيرة، بالإضافة إلى تصميم برمجية استعلام سريعة وشاملة عن نصوص الذخيرة.

وقد قدمت برمجية الاستعلام بذخيرة ليفاك إمكانات الاستعلام التالية:

أ) الاستعلام بالكلمة

حيث تكون نتيجة الاستعلام إظهار خصائص الكلمة بما في ذلك النطق القياسي ونطقها الدارج، بالإضافة إلى تقديم شرح لها باللغة الإنجليزية، مـع إظهار المعلومات الإحصائية الخاصة بها مثل عدد مرات ظهور هذه المفردة في كل منطقة لغوية من مناطق الذخيرة الخمس؛ وإدراج الكلمات المقابلـة، بمعنى إظهار مقابلات الكلمة في كل منطقة من مناطق الذخيرة.

ب) الاستعلام بالنص

السماح للمستخدم بالاستعلام باستخدام كلمة، أو جزء من كلمة، أو نوع من الأنواع النحوية للكلمات، أو النطق الصيني القياسي أو النطق الـدارج، بحيث تقترن أدوات الاستعلام تلك بمجموعة أخرى من الشروط المرتبطـة، وتكون نتيجة الاستعلام استخراج الجمل التي تنطبق عليها شروط الاستعلام.

ويمكن الاستعلام وفقًا لأربعة معايير هي: استعراض الكلمة المطلوب البحث عنها ومعها الكلمات المرافقة قبلها وبعدها بعدد معين من الكلمات، واستعراض التعبيرات اللغوية التي تحتوي كلمات معينة، واستعراض الجمل التي تحتوي كلمات معينة، بالإضافة إلى استعراض عناوين النصوص التي تحتوي كلمة معينة.

رابعًا: الذخيرة اللغوية لدراسات اللغة الصينية الحديثة

تم اعتماد الذخيرة اللغوية لدراسات اللغة الصينية الحديثة باعتبارها المشروع العلمي لقطاع العلوم الاجتماعية والإنسانية في الخطة الخمسية الثامنة لمجلس الدولة الصيني، وفي الوقت ذاته تم اعتمادها باعتبارها المشروع العلمي الرئيس لجامعة اللغات والثقافة في إطار الخطة الخمسية الثامنة للحكومة الصينية.

وكان الهدف البحثي من ذخيرة دراسات اللغة الصينية الحديثة هو تقديم قاعدة بحث لغوية أساسية ذات نطاق واسع يستخدمها العاملون في حقل معالجة المعلومات باللغة الصينية، وتعليم اللغة الصينية، والعاملون في حقل الدراسات اللغوية الصينية؛ وذلك من أجل دفع الدراسات اللغوية القائمة على اللغة الصينية إلى مزيد من التعمق[48].

١- اختيار المادة اللغوية وجمع العينات

تم تقسيم العمل في الذخيرة اللغوية للغة الصينية الحديثة إلى مستويين، المستوى الأول هو جمع ذخيرة لغوية خام يصل حجمها إلى ٢٠ مليون رمز

صيني، المستوى الثاني هو جمع ذخيرة لغوية سبق معالجتها تصل إلى ٢ مليون رمز صيني، أي ذخيرة لغوية تم عمل تمييز لحدود كلماتها وعمل ترميز لأنواع الكلمات بها.

وقد كان هناك مستويان لجمع المادة اللغوية للذخيرة.

المستوى الأول يتم عن طريق انتقاء مادة لغوية خام تصل إلى ٢٠ مليون رمز صيني من بين مادة يصل حجمها إلى ٦٠ مليون رمز.

وقد كانت قواعد اختيار العينات تتم على أساس الوضع في الاعتبار مدى اكتمال النص وطوله وغير ذلك من الاعتبارات، على سبيل المثال تُستَبعد النصوص التي يقل حجمها عن ألف رمز، وأيضًا النصوص غير المكتملة.

وقد اشتملت المادة الأساسية التي تم اختيار العينات منها، التي تصل إلى ٦٠ مليون رمز على النصوص الإخبارية للصين في الفترة من عام ١٩٩٢ إلى ١٩٩٣ بمعدل ٢٠ مليون رمز، والمراسلات الصحفية لشبكة الصين الجديدة شين خوا لعام ١٩٩٣ بحجم عشرة ملايين وخمسمائة ألف رمز، والنصوص الكاملة لجريدة الشعب اليومية لعام ١٩٩٤ بما يقرب من ٢٠ مليون رمز.

هذا بالإضافة إلى مختارات للكتب من مقتنيات المكتبات العامة الصينية في تخصص الأدب، ومختارات من اللغة الشفهية تصل إلى ٢ مليون وخمسمائة ألف رمز تم إدخالها إلى الحاسب الآلي عن طريق العنصر البشري.

وقد كانت القواعد التي اتُّبِعَت في اختيار ذخيرة العشرين مليون رمز من بين الحجم الإجمالي لتلك الذخيرة كما يلي:

(جريدة الشعب اليومية) (عام ١٩٩٤ بالكامل) ١٠,٠٠٠,٠٠٠ رمز.

(الأخبار الصينية) (١٩٩٢-١٩٩٣) ٥,٠٠٠,٠٠٠ رمز.

المؤلفات العلمية وغيرها ٢,٥٠٠,٠٠٠ رمز.

الأعمال الأدبية (عينات تم إدخالها يدويًا) ١,٥٠٠,٠٠٠ رمز.

(من بينها نصوص روائية تصل إلى مليون رمز، ونصوص نثرية تصل إلى ٣٠٠ ألف رمز، ونصوص في مجال الأدب التقريري تصل إلى ٢٠٠ ألف رمز).

مواد شفهية قياسية (عينات تم إدخالها يدويًا) مليون رمز.

(من بينها نصوص مسرحية تصل إلى ٦٠٠ ألف رمز، ومونولوجات تشمل مونولوجًا فرديًا، وخطبًا، وحوارات ثنائية، وقصصًا بما يصل إلى ٤٠٠ ألف رمز).

ومن الأرقام سالفة الذكر يمكننا ملاحظة أن المواد الصحفية في المستوى الأول قد وصلت إلى ٧٥%، والمؤلفات العلمية وصلت إلى ١٢,٥%، والأعمال الأدبية إلى ٧,٥%، والمواد الشفهية إلى ٥%.

أما المستوى الثاني فقد تم استبعاد الأخبار الصينية والمؤلفات العلمية منه؛ حيث تم انتقاؤه بشكل عشوائي وفقًا لنسب معينة قد سبق الاتفاق عليها في حدود مليوني رمز صيني من مادة لغوية يصل إجمالي حجمها إلى عشرة ملايين ومائتين وخمسين ألف رمز. وقد تم التفكير في (جريدة الشعب

اليومية) باعتبارها جريدة عامة غنية بـــالمواد اللغويـــة مـــن حيـــث تنـــوع الموضوعات، على عكس الجزء الباقي من المادة اللغوية، الذي يصل إلى ٧ ملايين وخمسمائة ألف رمز، التي تتميز بوحـــدة الموضـــوع والمـــضمون، بالإضافة إلى أن تلك الموضوعات يمثلها نسبة معينة في محتـــوى (جريـــدة الشعب اليومية). وعند تحديد نسب النصوص، تم مراعاة كلا من الموضوع والمضمون؛ حيث صُنِّفت النصوص وفقًا لهذين المعيارين. وعند التفكير في نسبة تمثيل كل نوع من النصوص داخل الذخيرة، تم الاهتمام بـــالجمع بـــين اكتمال النص ومبدأ النقطة المركزيـــة التـــي تـــدور حولهـــا موضـــوعات النصوص. على سبيل المثال، إذا تحدثنا عن الموضوع، فـــإن موضـــوعات السياسة، والاقتصاد، والأدب يكون لهما نسبة تمثيـــل أعلـــى مـــن التـــاريخ والجغرافيا والشئون العسكرية وغيرها من المجالات التـــي يكـــون التمثيـــل اللغوي لها ضعيفًا إلى حد ما لأنها تبعد عن الموضـــوع المركـــزي الـــذي تتمحور حوله نصوص الذخيرة. أما إذا تحدثنا من ناحيـــة المـــضمون فـــإن السرد والنقاش يكون التمثيل اللغوي لهما كبيرًا جدًا، أما الـــشرح والتطبيـــق فتكون النسبة على العكس من ذلك ضئيلة. وعلى ذلك تكون خطوات جمـــع العينات كما يلي:

١- بناء قاعدة بيانـــات لخـــصائص النـــصوص، تـــضم خـــصائص النصوص، وعنوان النص، وعدد الرموز المكونـــة لـــه، ومكـــان النشر، واسم الكتاب أو الجريدة، ودار النشر، بالإضافة إلى تـــاريخ النشر، وتصنيف الموضوع، وأسلوب الكتابة.

٢- تحديد طريقة توزيع المادة اللغوية.

٣- الاختيار العشوائي للعينات. وللتعرف على نتيجة اختيار العينـــات بالمرحلة الثانية انظر الجدول ٢-١٢ والجدول ٢-١٣ والجــدول ٢-١٤.

جدول (٢-١٢): توزيع الموضوعات داخل المادة التحريرية

الموضوع	النسبة المئوية	العدد (بالألف رمز)	عدد النصوص
سياسة وقانون	%١٥	٣٠	١٤٠
اقتصاد	%١٥	٣٠	١٢٩
أدب	%١٨,٧٥	٣٧,٥	١١٣
ثقافة وتعليم	%٧,٥	١٥	٨٣
حياة اجتماعية	%٧,٥	١٥	٨٠
علوم وتكنولوجيا	%٦	١٢	٦٩
رياضة	%٤	٨	١٩
جغرافيا وسياحة	%٢,٥	٥	٣٤
تاريخ وآثار	%١,٢٥	٢,٥	١٣
شئون عسكرية	%٢,٥	٥	٣٢

جدول (٢-١٣): توزيع الأساليب اللغوية داخل المادة التحريرية

أسلوب الكتابة	النسبة	العدد (بالألف رمز)
رواية	%١٠	٢٠
نثر	%١٠	٢٠
تقارير	%٢٥	٥٠
تقارير صحفية	%٥	١٠
مذكرات	%٢	٤
أبحاث علمية	%٩	١٨
مؤلفات نقدية	%١٤	٢٨
منتجات معرفية	%٠,٥	١
كتالوجات	%١	٢
ملخصات	%١	٢
نصوص وثائقية	%١	٢
قضايا تجارية	%١	٢
مراسلات	%٠,٥	١

جدول (٢-١٤): توزيع الأساليب اللغوية داخل المادة الشفهية

أسلوب الكتابة	العدد (بالألف رمز)	عدد النصوص
مسرحيات	٢٥,٥	٦٥
مونولوج فردي	٩	٢٨
مؤلفات نقدية	١,٢	٢
محاضرات وخطب	٣,٨	٣١
قصص	٣,٥	٢٦

وقد تم عمل تهيئة أولية للمادة اللغوية قبل معالجتها وذلك بالنسبة إلى المواد المُنْدَرِجة في إطار المستوى الثاني. ففي البداية تم تنقية جميع النصوص، على سبيل المثال تنقية النصوص المسرحية من الأجزاء التي تخرج عن نطاق الحوار، هذا بالإضافة إلى حذف المعلومات المتعلقة بالطبعات في العينات الأخرى من النصوص، ثم إضافة علامات تخص كل عينة؛ حيث تُعاد التسمية وفقًا لتصنيف الموضوع.

٢- معالجة الذخيرة

إذا أردنا عمل ترميز للذخائر اللغوية الصينية فلا بد أن تكون البداية بتمييز حدود الكلمات داخل الذخيرة. حتى يمكن جعل عملية التقسيم تحظى بموثوقية علمية، وضمان أن تكون عملية التقسيم موحدة، فقد تم وضع معيار على مستوى عالٍ من الدقة والوضوح، بالإضافة إلى إمكانية التطبيق عمليًّا. ففي البداية، تم الاستعانة بعدد مائتي ألف رمز من الرموز الصينية الممثِّلة للاستخدام اللغوي من حيث الشمول والانتشار، وتم البحث عن الرموز التي تُكَوِّن معها كلمات ثنائية وثلاثية المقطع، ثم عمل تحليل متعدد الاتجاهات لتلك الكلمات المركبة من عدد من الرموز، بما في ذلك طريقة التركيب

الداخلي، (هل يمكن للمكونات الداخلة في التركيب أن تُسْتَخْدَم منفردة، ونوع العلاقة بين العناصر المكونة للكلمة، ووظيفة تلك المكونات، وهل يمكن للعناصر الداخلية أن تتوسع، وما إلى ذلك)، والوظيفة العامة للكلمة، وطبيعة التركيب الدلالي، وتركيب الوحدات الصوتية، والعناصر الأسلوبية وغير ذلك من المجالات. وعلى هذا الأساس يتم التأكد من فاعلية المنهجية والنظرية المتبعة في عملية تمييز حدود الكلمات، إلى أن نصل في النهاية إلى تحديد قواعد تقسيم النصوص الصينية إلى كلمات. وتتضح القواعد المُتَّبَعة في تقسيم النصوص إلى كلمات من خلال الجوانب الثلاثة التالية:

١- الكلمة لها حدود دلالية ونحوية.

٢- الحدود التي يتم تقسيم الكلمات على أساسها ليست مطلقة.

٣- ينبغي مراعاة التمييز بين المستويات المختلفة للمادة اللغوية. وفي النهاية يتم التوصل إلى الحكم عما إذا كانت الوحدة اللغوية المفردة تدخل ضمن إطار كلمة مركبة أم أنها تمثل وحدة مستقلة بذاتها.

وعلى هذا الأساس، يتم عمل ترميز نحوي للكلمات المكوِّنة للمستوى الثاني من ذخيرة دراسات اللغة الصينية الحديثة. وبعد الاطلاع على نتائج الدراسات في حقل اللسانيات الصينية، ونتائج الأبحاث المتعلقة بترميز الكلمات الصينية في حقل اللغويات الحاسوبية الصينية، تم الاستعانة بقاعدتين أساسيتين في عملية تحديد نوع الكلمات الصينية، الأولى تحديد نوع الكلمة بصورة كاملة بناءً على وظيفتها النحوية داخل الجملة. والثانية تحديد نوع الكلمة بشكل تقريبي من حيث درجة الدقة.

ومن أجل ذلك، تم استخدام نظام متعدد المستويات لتحديد نوع الكلمة في نصوص الذخيرة.

على سبيل المثال، أول مستوى من مستويات ترميز الاسم هو استخدام الرمز "**n**"، ثم يتم تصنيف الاسم بعد ذلك إلى خمسة أنواع فرعية هي اسم العلم، والاسم العام، واسم الزمان، واسم المكان، واسم الموقع.

وقد تم تصميم مجموعة من الرموز الخاصة لترميز عدد ٨٥ نوعًا من أنواع الكلمات الصينية.

ومن أجل إتمام عملية ترميز المادة اللغوية استُخدِمَت أدوات الترميز وتمييز الكلمات المتاحة في نظام **CCID**. وبعد انتهاء الترميز الآلي، تم عمل مراجعة بالمجهود البشري لنتيجة ترميز مليوني رمز صيني. ومن أجل تهيئة ذخيرة الدراسات الصينية الحديثة لخدمة الباحثين في حقل اللغويات الصينية، تم تطوير واجهة تطبيقية بها إمكانات بناء ذخيرة، واستعلام، وإحصاء، وما إلى ذلك.

باستخدام المستوى الأول من الذخيرة يمكن عمل استعلام عن سلاسل الرموز الواردة في الذخيرة، وباستخدام المستوى الثاني من الذخيرة يمكن البحث باستخدام الكلمة أو نوع الكلمة أو المعلومات الإحصائية، على سبيل المثال الاستعلام عن الجمل التي تحتوي نوعًا معينًا من الكلمات، ويمكن أيضًا الاستعلام عن نوع معين من التعبيرات اللغوية، بالإضافة إلى عرض النتيجة بأكثر من صيغة.

خامسًا: الذخيرة اللغوية الصينية ذات المعالجة الدقيقة

الذخيرة اللغوية الصينية المُعَالَجة بدقة هي مشروع مركزي تم تمويله من الحكومة الصينية في مجال العلوم الطبيعية، وعنوانه الفرعي هو "نظريات البحث في الذخائر اللغوية ومنهجياته وأدواته"، والهدف من هذا العنوان الفرعي هو إنشاء سلسلة من الذخائر اللغوية الصينية التي تمت معالجتها جيدًا، والهدف من هذه السلسلة من الذخائر هو التوصل إلى إصدار عدة مؤلفات تحمل العناوين الآتية: (توحيد معياري لطرق تمييز الكلمات في ذخائر اللغة الصينية الحديثة)، و(توحيد معياري لترميز الكلمات في اللغة الصينية الحديثة)، و(مبادئ اختيار المادة اللغوية وتوزيعها داخل الذخيرة). وفي النهاية، تم الحصول على ذخيرة متوازنة مكونة من مليوني رمز صيني (لا تشمل علامات الترقيم)، وتم إجراء عمليات تمييز لحدود الكلمات بداخلها، وعمل ترميز لأنواع الكلمات وبعض المعلومات النحوية، هذا بالإضافة إلى ذخيرة خام مُعْتَمَدة يصل حجمها إلى ١٢٠ مليون رمز صيني.

والنتيجة التي حققتها هذه الذخيرة يمكن أن تكون واجهة تطبيقية يُعْتَمَد عليها وتتمتع بموثوقية في الدراسات المتعلقة بالنحو والصرف في اللغة الصينية الحديثة، ومن ثَمَّ، يكون لها مغزى مهمّ في الدراسات العلمية القائمة على تعليم اللغة الصينية، واللسانيات الصينية، ومعالجة المعلومات باللغة الصينية.

١- اختيار المادة اللغوية للذخيرة

أ) المبادئ المُتَّبَعة في اختيار المادة اللغوية

تم الالتزام بالمبادئ التالية في أثناء اختيار المادة اللغوية

أن تكون المادة التي يتم اختيارها تنتمي إلى فترة تسعينيات القرن العشرين (مع جزء صغير في فترة الثمانينيات) بحيث يمكنها أن تعكس ملامح اللغة الصينية المعاصرة.

أن تكون وحدة اختيار المادة اللغوية هي النصوص الكاملة؛ وذلك من أجل الحفاظ على معلومات النص كاملة.

أن يكون المعيار الأساسي في جمع المادة هو النصوص موزعة على الأساليب اللغوية المختلفة، والمعيار الفرعي هو التوزيع حسب المجالات؛ حيث إن الأسلوب اللغوي يسبق المجال اللغوي من حيث الأهمية.

أن تركز المادة اللغوية التي تم جمعها على المستوى التحريري، بالإضافة إلى المواد الشفهية التي يمكن تدوينها في كتب، مثل نصوص المسرح، والتسجيل الصوتي للحوارات، والتسجيل الصوتي للمحاضرات وما إلى ذلك. وقد تم تمثيل المواد اللغوية الأدبية بنسبة أكبر، حتى يمكن زيادة مقدار اللغة الشفهية ولغة الاستعمال اليومي بصورة مناسبة؛ وذلك من أجل تجنب أن تتجه الذخيرة إلى الاقتصار على اللغة التحريرية البحتة.

ألا تتضمن المواد اللغوية مطبوعات من مناطق هونج كونج، ومكاو، وتايوان.

ب) تفاصيل توزيع المادة اللغوية

تم تقسيم المادة اللغوية حسب الأسلوب اللغوي إلى أربعة أنواع من النصوص هي: نصوص أدبية، ونصوص صحفية، ونصوص علمية، ونصوص تطبيقية، وقد وصل حجم تلك النصوص إلى مليوني رمز صيني.

وقد تم توزيع كل نوع من تلك النصوص كما يلي:

الأدب:

١) الروايات (ويشمل القصص العامة، والقصص العاطفية، وقصص الخيال العلمي، والقصص البوليسية).

٢) النثر (ويشمل نصوصًا متفرقة، ومقالات قصيرة متنوعة وما إلى ذلك).

٣) المذكرات (وتشمل السير الذاتية).

٤) أدب التقارير.

٥) المسرحيات (وتشمل الحوارات، والخطب).

الأخبار:

١) التقارير الإخبارية (وتشمل السياسة، والاقتصاد، والشئون العسكرية، والصناعة، والزراعة، والتجارة، والعلوم والتكنولوجيا، والرياضة وغير ذلك من المجالات).

٢) النقد الاجتماعي والتعليقات.

٣) الحياة الاجتماعية والترفيه (وتشمل السياحة، والطهو، والأزياء، والفنون، والعادات وغيرها).

المؤلفات العلمية (وتشمل العلوم الاجتماعية والإنسانيات).

النصوص الوظيفية (بما في ذلك الإعلانات، والإخطارات، والمراسلات، والتقارير، والعقود، ومذكرات التفاهم، وكتالوجات المنتجات، وغير ذلك من المطبوعات غير الرسمية).

للاطلاع على التوزيع التفصيلي للمادة اللغوية انظر الجدولين ٢-١٥، و٢-١٦.

جدول (٢-١٥): معلومات إحصائية عن توزيع المادة اللغوية بالذخيرة الصينية ذات المعالجة الدقيقة

التصنيف	عدد النصوص	عدد الرموز	النسبة	عدد علامات الترقيم	عدد الكلمات	النسبة
الأدب	٢٩٥	٨٨٠٠٥٧	%٤٤	١٤٨٤٥٣	٧٦٠٣٣٧	%٤٨
الأخبار	٣٧٦	٦٠٠٤٩٠	%٣٠	٨٦١٦٣	٤٣٨٠٩٥	%٢٨
العلوم	٢٩	٤٠٢٦٢٣	%٢٠	٥٢٨٢٣	٢٧٨٧٢٨	%١٨
النصوص الوظيفية	٢٥٨	١١٩٤٨٨	%٦	٢٨٧٢٧	٩١٩٢٩	%٦
الإجمالي	٩٥٨	٢٠٠٢٦٥٨	%١٠٠	٣١٦١١٦	١٥٦٩٠٨٩	%١٠٠

جدول (٢-١٦): معلومات إحصائية عن توزيع النصوص الأدبية بالذخيرة

التصنيف	عدد النصوص	عدد الرموز	النسبة	عدد علامات الترقيم	عدد الكلمات
الروايات	١٩٩	٦٤٨٧٩٦	%٣٢,٥	١١٢٧٤٩	٥٦٦٧٣
النثر	٣٧	٨٠٠٦٧	%٤	١٠٣٤٧	٦٥٤٥٣
المذكرات	٢٩	٥٠٤٠١	%٢,٥	٦٩٠٨	٣٨٣٣٨
الأدب التقريري	١٣	٥٠٠١٩	%٢,٥	٨٢٢٥	٤٠٣٨٦
النصوص المسرحية	١٧	٥٠٧٧٤	%٢,٥	١٠٢٢٤	٤٩٤٣٠
الإجمالي	٢٩٥	٨٨٠٠٥٧	%٤٤	١٤٨٤٥٣	٧٦٠٣٣٧

٢- نطاق تمييز الكلمات

عند وضع حدود الكلمات الصينية، تم مراعاة النقطتين التاليتين:

أ- الاستقرار على منهجية وضع حدود الكلمات على أساس الأبحاث واسعة النطاق التي أُجريت على تمييز الكلمات الصينية.

١- الاستفادة من النتائج السابقة في هذا المجال، وعلى الأخص الاستعانة بنتائج الأبحاث في مجال اللغويات الحاسوبية واللغويات الصينية، والسعي قدر الإمكان إلى التوافق مع المعايير الصينية القومية التي تم نشرها من قبل فيما يتعلق بهذا الأمر. ولكن مع مراعاة أن يكون لنظام التمييز الجديد سماته الخاصة.

٢- عمل دراسة متخصصة عن مجموعة الرموز الخاصة التي تحمل لبسًا لغويًّا في عملية التمييز. وقد تمت مراجعة النتائج التي تصل إلى مليون وخمسمائة ألف رمز صيني من خلال العنصر البشري عن طريق سبعة طلاب للدراسات العليا؛ حيث تم رصد سبعمائة وعشر ألف رمز صيني من بين العدد الإجمالي للرموز لم يحدث لها تمييز بشكل مُوَحَّد. وقد تم عمل تحليل يدوي لهذه الوحدات اللغوية التي الْتَبَسَت على الآلة. ومن نتيجة الأبحاث نكتشف أن عدم وضوح الحدود الفاصلة بين الكلمات والتعبيرات اللغوية يتركز أساسًا في عدم وضوح الحدود بين الكلمات ثنائية وثلاثية المقطع في اللغة الصينية، أما من ناحية الوظيفة النحوية فيتركز عدم الوضوح بشكل أساسي في التراكيب الاسمية والفعلية.

٣- اتباع منهجية تطبيق عمليات التمييز بالتوازي مع إجراء التصحيحات على النتائج، بغرض التوصل إلى معيار دقيق لتمييز الكلمات يتمتع بتغطية أكثر شمولاً. وقد كان تحديد المعايير القياسية لتمييز الكلمات لا يتم بصورة صماء، بل اتُّبِعَت المعايير على أساس التوزيع الحقيقي للمادة اللغوية داخل البيئة اللغوية الواقعية.

ب- تبني الإجراءات الضرورية لضمان تمييز قياسي وموحد لتمييز الكلمات.

١- اتخاذ قرارات مُلْزِمة بشأن الوحدات اللغوية التي تحمل لبسًا لغويًّا لمدة طويلة بين المتخصصين. على سبيل المثال، التركيب الإضافي المكوَّن من رمزين إلى أربع رموز صينية بالصيغة "اسم + اسم"، و"فعل + اسم" إذا كان طول أحد عناصره رمزًا واحدًا لا يتم تقطيعه إلى وحدات أقل.

٢- العديد من العناصر الدقيقة المكوِّنة للمعيار تُستَخدم خصيصى باعتبارها كلمة أو عدة كلمات. عل سبيل المثال الرمز "半" الذي يعني "نصف" عندما يعبر عن عدد مستقل يتم تمييزه على أساس أنه كلمة، مثل:

半/斤[١]

半/斤/一

(١) كلمة "斤" وحدة قياس صينية للوزن تساوي ٥٠٠ جرام. (المترجم)

بمعنى:

نصف / نصف كيلو (ربع كيلو)

واحد / نصف كيلو / نصف (كليو إلا ربع)

إلا أن الوحدات التالية التي تحتوى الرمز "半" يتم تمييزها على أنها كلمات متعددة المقطع، ولا يمكن تقسيمها إلى وحدات أصغر خلافًا للمثالين السابقين، مثل:

一半儿、多半儿、两半儿、大半儿、一多半儿、一大半儿。

٣- معايير ترميز تصنيف الكلمات

إن المعايير التي تم اعتمادها لترميز تصنيف الكلمات تشمل عدد ١١٩ علامة، من بينها عدد ٩٥ علامة لترميز تصنيف الكلمات، وعدد ٢٤ علامة لترميز علامات الترقيم. وقد تم اتباع نظام متعدد المستويات لاستخدام العلامات الخمس والتسعين لترميز تصنيف الكلمات، على أن يكون أكبر قدر من المستويات هو ثلاثة مستويات، مثل: "npf"، فالعلامة "n" تمثل المستوى الأعلى، وتشير إلى الاسم، أما العلامة "p" فتشير إلى المستوى الأوسط، وتعني تصنيفًا فرعيًا للاسم وهو اسم العلم، أما العلامة "f" فتشير إلى المستوى الثالث، وتعبر عن اسم علم أجنبي، يختلف عن أسماء الأعلام اليابانية والكورية والصينية التي تتبع الأسلوب الصيني في تسمية الأعلام الشخصية.

ويستعين المستوى الأعلى من العلامات عدد ٢٢ علامة، منها العلامة "n" التي تشير إلى الاسم، والعلامة "v" التي تشير إلى الفعل، والعلامة "a"

التي تشير إلى الصفة، والعلامة "z" التي تشير إلى الكلمات المعبرة عن الحالة، والعلامة "b" التي تشير إلى التمييز، والعلامة "t" التي تشير إلى اسم الزمان، والعلامة "s" التي تشير إلى اسم المكان، والعلامة "f" التي تشير إلى اسم الموضع، والعلامة "m" التي تشير إلى الأرقام، والعلامة "q" التي تشير إلى الكلمات الكمية، والعلامة "d" التي تشير إلى الظرف، والعلامة "r" التي تشير إلى الضمائر، والعلامة "o" التي تشير إلى الكلمات الصوتية، والعلامة "e" التي تشير إلى كلمات التعجب، والعلامة "c" التي تشير إلى حروف العطف، والعلامة "p" التي تشير إلى حروف الجر، والعلامة "u" التي تشير إلى الكلمات المساعدة، والعلامة "y" التي تشير إلى كلمات اللهجة، والعلامة "I" التي تشير إلى الكلمات الاعتراضية، والعلامة "i" التي تشير إلى الأمثال، والعلامة "k" التي تشير إلى اللواحق، والعلامة "x" التي تشير إلى الأرقام العربية والحروف الإنجليزية.

وقد تميزت المعايير المُسْتَخدَمة في ترميز تصنيف الكلمات الصينية بالخصائص التالية:

أ) الاهتمام الشديد بالنقاط الصعبة في ترميز تصنيف الكلمات. فعندما يتم وضع رمز لنوع الكلمة داخل الذخيرة يمكن بصورة عامة مواجهة الحالات التالية:

١- الكلمات التي تتمتع بثبوت المعنى وأحادية الوظيفة النحوية لا يحدث تجاهها ازدواج في وضع الترميز، ويصبح من السهل الاستدلال على الفئة التي تنتمي إليها ووضع الرمز الخاص بها.

٢- الكلمات ذات الازدواجية النحوية التي لها وظائف نحوية مختلفة؛ التي تمثل اختلافات دلالية كبيرة (بما في ذلك الكلمات المتفقة في الشكل) يمكن أن ينطبق عليها معياران أو أكثر من معايير الترميز.

٣- هناك قدر من الكلمات على الرغم من عدم اتفاقها في الوظيفة النحوية، فإن هناك اختلافًا حول مدى اعتبارها كلمات ثنائية التصنيف.

٤- يمكن إدراج الكلمة تحت تصنيفين أو أكثر من وجهة النظر النحوية أو الدلالية؛ ولكن من الناحية العملية لا يمكن أن يتم إدراجها تحت تصنيفين؛ مثل الصفات والأفعال اللازمة.

٥- الكلمات التي لا يوجد اتفاق حول وظيفتها النحوية، ومن الصعب العثور على تصنيف مناسب لها.

٦- فئات أخرى، بما في ذلك المشكلات التي تنشأ من تمييز الكلمات، والاستخدام المؤقت للكلمات وما إلى ذلك. وتتركز الصعوبات الخاصة بترميز أنواع الكلمات بشكل أساسي في النقاط الأربع الأخيرة؛ مثل الصفات وحدود الأفعال اللازمة، والتمييز بين كل من الكلمات والظروف، وبين كل من الأسماء، والأفعال، والصفات، وغير ذلك من مشاكل تمييز الأنواع النحوية للكلمات. لأنه عند وضع المعايير القياسية للترميز يتم الانتباه بشدة إلى هذه الأنواع.

ب) هناك جزء من الكلمات تم عمل ترميز متعدد المستويات لها، الأمر الذي يسهل عملية تقديم المعلومات الخاصة بترميز تصنيف

الكلمات، تلك المتعلقة بعدد من الدراسات الخاصة، بالإضافة إلى إمكانية إجراء بحوث متخصصة عن بعض أنواع الكلمات التي لها تمثيل ضعيف داخل الذخيرة.

ج) البحث الدقيق عن أسماء الأعلام، وبصورة خاصة أسماء الأشخاص؛ حيث يتم تمييزها بصورة دقيقة، وتقديم معلومات وافية وغنية عن الاسم وطرق التعرف عليه.

د) الترميز الجزئي للمعلومات النحوية. تتركز تلك المعلومات بصورة أساسية في الجزء المتعلق بالفعل داخل الجملة (فالصفة التي يكون لها موصوف يتم ترميزها على أنها فعل). وهكذا يمكن عمل دراسات متخصصة متعلقة بالفعل مثل بعض التعبيرات اللغوية التي يمكن التعامل معها من زاوية التعدي أو اللزوم، ويمكن أيضًا عمل بعض الدراسات المتعلقة ببعض الأنماط الخاصة للجمل.

هـ) إعطاء علامات خاصة لبعض الكلمات التي لها نسبة تكرار مرتفعة، من أجل تسهيل عمل دراسات متخصصة حولها.

٤- ضمان مستوى الجودة في معالجة الذخيرة الصينية

إن مرحلة معالجة الذخيرة من حيث تمييز حدود الكلمات وترميزها نحويًا يتم تنفيذها آليًا؛ ولكن على أساس أن يتم عمل مراجعة بالطاقة البشرية لنتيجة التمييز والترميز؛ حيث تُعتَبَر المراجعة البشرية للنتيجة من أهم مراحل العمل في الذخائر اللغوية بصورة عامة.

إن المراجعة البشرية لنتيجة معالجة الذخائر اللغوية يتضمن مجالين أساسيين هما: عملية تمييز الكلمات وعملية تحديد أنواع الكلمات. فأكثر المشكلات ظهورًا في مرحلة تمييز الكلمات في الذخيرة هي عدم التوحيد في عملية تمييز الكلمات التي تتكون من مقاطع متشابهة، أما أكثر المشكلات ظهورًا في عملية ترميز الأنواع النحوية الكلمات فتتمثل في أن يكون هناك تشابه تام بين الشكل والوظيفة النحوية ومعنى الكلمة، ويؤدي ذلك إلى عدم توحيد في عملية الترميز.

وهذا النوع من المشكلات يتم مراجعته على مستويات منفصلة وفقًا لطبيعة كل مشكلة، بالإضافة إلى أن كل مرحلة يتم فيها اتباع وسائل مختلفة بهدف الوصول إلى قرار موحَّد بشأنها.

١- مراجعة تمييز الكلمات

تنقسم عملية مراجعة تمييز الكلمات إلى مرحلتين: الأولى يتم فيها قراءة تتبعية للنص، يصاحبها قراءة تتبعية للمفردات التي تم تمييزها، والخطوة الثانية يتم فيها كتابة برمجية استعلام عن الكلمات المتفقَة من حيث التقسيم، بغرض استخراج جميع الحالات التي تمثل لبسًا في عملية التمييز، واستخدام ذلك في عمل مراجعة تفاعلية مع قائمة المفردات المستخرجة.

٢- مراجعة الترميز النحوي للكلمات

تنقسم عملية مراجعة الترميز النحوي للكلمات أيضًا إلى مرحلتين، المرحلة الأولى يُتَّبَع فيها عملية القراءة التتبعية للنص بالتوازي مع القراءة

التتبعية للكلمات وترميزها. أما المرحلة الثانية فبغرض التغلب على مشكلة التوحيد في معايير تمييز الكلمات في أثناء مراجعة نتيجة تمييز الكلمات، فإن فريق إنشاء الذخيرة كان في حاجة إلى تكوين قائمة تضم جميع الكلمات المكوِّنة للذخيرة وعلامات الترميز المستخدمة؛ وذلك من أجل عمل مراجعة تفاعيلة لمراجعة نتيجة ترميز أنواع الكلمات. وبذلك يمكن تجنب وضع علامات مختلفة للوحدات اللغوية نفسها.

وهذه الذخيرة من المنتظر رفعها على شبكة المعلومات لخدمة أكبر عدد من الباحثين.

الباب الثالث

معالجة الذخائر اللغوية وتقنيات إدارتها

الفصل الأول

الاستعلام داخل الذخائر اللغوية وتطبيقاته

بعد الانتهاء من بناء الذخيرة اللغوية، يتم عرضها للاستخدام من قِبَل جميع الباحثين؛ بغرض تمكينهم من التعامل مع مضمون الذخيرة، وإجراء عمليات التحليل للظواهر اللغوية الواقعية الموجودة في مادتها اللغوية. لذلك فإن أبسط ما يمكن أن تُقدمه الذخيرة اللغوية للمستخدم هو أدوات الاستعلام عن المواد اللغوية داخل الذخيرة.

أولاً: الاستعلام الإحصائي السياقي عن الكلمات

عادة ما تلجأ الذخائر اللغوية المُمَيْكَنة إلى أسلوب الاستعلام الإحصائي السياقي عن الكلمات "**concordance**" لتقديم المعلومات الإحصائية المتعلقة بالسياقات التي تظهر فيها كلمة معينة داخل متن الذخيرة. وتُسَجِّل البرامج المختصة بالاستعلام الإحصائي عن سياقات الكلمات موقع الكلمة موضع البحث في كل مرة ظهرت فيها داخل الذخيرة، وبناءً على ذلك يمكن تقديم المعلومات السياقية المتعلقة بتلك الكلمة. وهذه المعلومات يمكن أن تظهر مباشرة على شاشة الحاسب أو يتم حفظها في ملف معين. وهذا الملف الذي يتم حفظه يُطْلَق عليه اسم ملف الإحصاء السياقي للكلمات "**concordance file**"[49].

وقَبْل عمل استعلام إحصائي عن سياقات الكلمات، تكون هناك حاجة لبناء فهرس لكل كلمة من كلمات الذخيرة، يُسَجَّل في هذا الفهرس موقع هذه الكلمة داخل النص في كل مرة من مرات ورودها، ثم يتم إعادة فهرسة ملف البيانات هذا وفقًا للأنماط الجديدة التي يحتويها؛ وذلك من أجل تسهيل عملية البحث في داخله؛ كأن يتم بناء الجدول المفهرس للذخيرة وفقًا للترتيب الأبجدي لكلماتها. وبذلك يُمْكِننا في أثناء عمل الاستعلام عن سياقات الكلمات، الحصول بسهولة عن السياقات التي وردت فيها الكلمة موضع البحث.

إن أبسط ما يُقَدِّمه البحث الإحصائي عن سياق الكلمة هو الاستعلام بموضع الكلمة؛ حيث يُقَدِّم هذا الاستعلام عرضًا مُفَهْرَسًا لموقع الكلمة محل البحث في كل مرة وردت فيها داخل النص، بالإضافة إلى إمكانية تقديم إحصائية عن معدل ظهور هذه الكلمة داخل الذخيرة بأكملها. وهناك نوع آخر من أنواع الاستعلام السياقي هو الاستعلام عن السياقات التي ترد فيها كلمةٌ ما داخل الذخيرة؛ حيث يُقَدِّم السياقات التي تَرِد فيها الكلمة موضع البحث.

١ - الاستعلامْ عن كلمة مفتاحية داخل السياق

أهم نوع من أنواع الاستعلامات وأكثرها شيوعًا هو الاستعلام عن كلمة مفتاحية داخل السياق "Key Word in context" والذي يطلق عليه اختصارًا اسم KWIC. وفي ذلك الاستعلام تظهر الكلمة المُستَعلَم عنها في منتصف كل سطر، وقبلها وبعدها مسافة، ويلي كل مسافة منهما سياق نصي بعدد من الكلمات يمكن التحكم في طوله. وفي هذا الأسلوب من أساليب

العرض تتراص الكلمات موضع البحث في كل سطر مُكَوِّنة عمودًا رأسيًا ظاهرًا للعيان. ويُظهِر الشكل ٣-١ استعلامًا عن كلمة "is" يسبقها ويليها سياق من أربع كلمات.

of activity and communication	is	only one of them
communication where the activity	is	halted in time if
whole process the activity	is	obvios enough the nervous
reader in his armchair	is	making continuous fast and
radio listener his brain	is	highly active if he
human communication through language	is	only a small sub-section

شكل (٣-١): نتيجة البحث السياقي KWIC عن كلمة "is" في اللغة الإنجليزية

كما يمكن تعديل طول السياق المصاحب للكلمة من جهة اليسار وَجِهَة اليمين حسب الحاجة؛ حيث يمكن زيادة أو نقصان عدد ٦ كلمات من جهة اليمين أو اليسار. وهناك بعض أدوات الاستعلام الحاسوبية يمكنها توسيع ما يُعْرَض في نتيجة الاستعلام لتستخرج كامل الجملة أو الفقرة التي وردت فيها الكلمة موضع الاستعلام.

٢- الترتيب المُفَهْرَس للأسطر

هذا النوع من الاستعلام له عدة أشكال، الشكل الأكثر شيوعًا هو: ترتيب السطر المُستَعلَم عنه وفقًا لترتيب ظهوره داخل الذخيرة، أو ترتيب السطر المُستَعلَم عنه وفقًا للترتيب الأبجدي لكلماته. على سبيل المثال، ترتيب نتيجة البحث حسب الترتيب الأبجدي للكلمة التي تأتي يمين الكلمة المفتاحية. وهذا النوع من الترتيب يَعرِض الكلمات التي تستدعيها الكلمة المفتاحية؛ التي تُعتَبر في تلك الحالة هي الكلمة المركزية التي تدور حولها عملية البحث. وهناك طريقة أخرى للترتيب هي فهرسة الأسطر حسب الترتيب الأبجدي لآخر حرف في الكلمة المجاورة للكلمة موضع البحث من جهة اليسار. وما إن تكون الكلمة المفتاحية فعلاً من حيث التصنيف النحوي، حتى يمكن لهذه الطريقة في البحث أن تعثر وبسهولة على الفاعل في كل جملة؛ ومن ثَمَّ يمكن تقديم بعض الخيوط التي من شأنها أن تساعد في التعرف على عملية الاقتران الإسنادي وعناوين النصوص. كما يمكن الحصول على الفهرسة حسب الكلمات الواقعة يمين الكلمة موضع البحث ويسارها؛ حيث يتم وضع الكلمات المصاحبة للكلمة موضع البحث معًا حسب أكثرها تواترًا، وهذه الطريقة من طرق إظهار نتيجة الاستعلام لها استخدامات عظيمة في الدراسات التي تبحث في قواعد التصاحب اللغوي بين الكلمات.

وبالنسبة إلى الكلمات ذات معدل التكرار الأعلى داخل الذخيرة، يمكننا أيضًا اللجوء إلى أسلوب الاستعلام من خلال العينات لتقليل نطاق العينات

المستخدمة في الاستعلام. على سبيل المثال، عند الاستعلام عن كلمة معينة يمكننا ضبط الرقم الاسترشادي للاستعلام على ١٠؛ بحيث يتم الاستعلام مرة كل عشرة أسطر، وبذلك يمكن أن نقلِّص نتيجة الاستعلام بنسبة العُشر عن النتيجة نفسها إذا أُجْرِيَت دون هذا التدخل.

وفي اللغة الإنجليزية، يمكننا استخدام علامة النجمة * للاستعلام عن الكلمات التي لها علاقة بالأصل الاشتقاقي لكلمة معينة. مثل البحث بالصيغة **photo*** فتكون نتيجة الاستعلام هي: **photo, photograph, photographer, photography** وغيرها من الكلمات المُولَّدة من الكلمة الأساسية موضع البحث بالإضافة إلى الأسطر التي احتوت سياقات لها. كما يمكن أيضًا الاستعلام باستخدام اللواحق النحوية والتعرف على بعض الكلمات الخاصة وخصائصها النحوية. على سبيل المثال، استخدام صيغة الاستعلام "**ng***" لسرد جميع الكلمات التي تحتوي اللاحقة "**ing**"، واستخدام الصيغة "**?***" للاستعلام عن الجمل الاستفهامية.

استخدام المنطق المركب وصيغة "**with/n**"، للاستعلام عن السياقات التي تفصلها كلمات معينة لها علاقة بالكلمة موضع البحث سواء كانت تعبيرات لغوية أو جزء من جملة.

ثانيًا: تطبيقات الاستعلام بالكلمة داخل الذخائر اللغوية

تتيح الاستعانة بالقدرات الهائلة للحاسب الآلي في الحساب ومعالجة المعلومات، تُتيح للمستخدم الاستفادة من برمجيات الاستعلام عن الكلمات في

استخراج المحتويات التي يرغب في الحصول عليها من الذخيرة، بالإضافة إلى مراقبة تلك الكلمات داخل البيئة اللغوية المصاحبة لها، أو إجراء دراسات مقارنة عن الظواهر اللغوية المتعلقة بها. إن هذا النوع من التطبيقات القائمة على الذخائر اللغوية له قيمة هائلة، ويتمتع بقدرة تطبيقية واسعة النطاق تزداد يومًا بعد يوم في مجالات الأدب وعلم اللغة، وتعليم اللغات وأنظمة معالجة اللغات الطبيعية، بالإضافة إلى ما حققه بالفعل من نتائج مذهلة حتى الآن. وفيما يلي قائمة بمجالات التطبيق سالفة الذكر[23].

١- التطبيقات في مجال الأدب

يتمثل ذلك في بناء ذخائر لغوية للأعمال الأدبية المشهورة، وإنشاء ملفات مفهرسة قائمة على تلك الذخائر. وتُستخدم ملفات الفهرسة على نطاق واسع في مجال الدراسات الأدبية؛ حيث تُقدم بيانات مثالية للتعرف على خصائص أعمال أحد الأدباء أو الشعراء، وطرق استخدامه للكلمات وأسلوبه اللغوي.

إن إنشاء ملف مفهرس عن الأعمال الأدبية وإتاحته للاستخدام من قِبَل علماء اللغة يُمَكِّنهُم من دراسة خصائص أسلوب أدبي معين في الكتابة؛ فعلى سبيل المثال، من خلال إنشاء ملف مفهرس للأعمال الأدبية لشكسبير، يمكننا أن ندرس الصيغ المختلفة لضمير المخاطَب في اللغة الإنجليزية مثل (**ye, you, your, yours, thou, thee, thy, thine**) ومجالات استخدام كل صيغة، ومن ثَم يمكن طرح بعض الفروض الخاصة بعلم اللغة الاجتماعي وقياسها، على غرار: مَن المتحدث؟ ومَن المتحدَّث إليه؟ وما ظروف الحديث وملابساته؟ وما إلى ذلك. وعلى سبيل المثال أيضًا، يُمكِننا عن طريق

الاستعلام بالكلمات أن ندرس ظروف استخدام بعض الكلمات للتعرف على اتجاهات تطور الاستخدام اللغوي على مدى مائتي عام من خلال دراسة لغة أعمال الشاعر الألماني جوته. فمثلا كلمة "knable" كانت تُستَخدَم على نطاق واسع في شعر جوته (Goethe) إذا ما قورن ذلك باستخدامها في الأعمال الحديثة. وجميع الأنماط اللغوية الأخرى يمكن التعامل معها بمثل تلك الطريقة. كما يمكن عمل دراسات مقارِنة للأعمال الكلاسيكية مع الأعمال الأدبية الحديثة، وعمل تحليل كمي لمثل تلك التغيرات التي تحدث للغة مع مرور الزمن. ويمكن أيضًا استخدام الاستعلام بالكلمة لعمل دراسات عن الكلمات وأنواعها والتراكيب النحوية داخل الأعمال الأدبية من خلال عدة زوايا للتحليل؛ وذلك بغرض التعرف على أوجه الاختلاف والتشابه بين أعمال الكُتَّاب في مراحلهم الإبداعية الأولى والمتأخرة، فضلاً عن دراسة أوجه الاختلاف والتشابه بين أساليب عددٍ من الكتاب، بالإضافة إلى تقديم أدلة قوية للحكم على مدى انتماء بعض الأعمال الأدبية لفترات معينة من فترات الإبداع الأدبي.

٢- التطبيقات في مجال الدراسات اللغوية

أشار ليتش (Leech:1992) إلى أن الأداة الأكثر بساطة وفاعلية والأكثر استخدامًا والقائمة على أساس حاسوبي هي الاستعلام عن الكلمات. وتتمثل التطبيقات النموذجية لذلك في:

١- الدراسات في علم المفردات: حيث يتم التوصل إلى مغزى الكلمات وما تحمله من معنى في سياق محدد، بالإضافة إلى طرق استخدام تلك الكلمات من حيث القواعد، وطبيعة الأسلوب، والسياق وما إلى ذلك.

٢- الدراسات النحوية التصنيفية: تمييز وتصنيف الأمثلة النموذجية المستخدمة في بعض المجالات من حيث النحو، والدلالة، والسياق، والأسلوبية. والذخائر اللغوية المُرَمَّزة صرفيًا ونحويًا هي أكثر الذخائر ملائمة لهذا الغرض.

٣- في مجال تعليم اللغات: في أثناء تأليف المواد الدراسية لتعليم اللغة الإنجليزية، قام بيبر (**Biber**) وآخرون عام ١٩٩٤ بعمل دراسة عن الأساليب البلاغية التي تتكون من الاسم + صفة في اللغة الإنجليزية[50]؛ حيث قاموا بعمل دراسة مسحية عن شرح هذا التركيب في عدد من المؤلفات في قواعد اللغة الإنجليزية في فترة ثمانينيات القرن العشرين، واكتشفوا أن غالبية هذه المؤلفات النحوية تهتم بهذه العلاقة بصورة ملحوظة، لدرجة أن شرح هذه القاعدة قد وصل إلى ٦٠ صفحة. أما النصوص التي تتناول التركيبات البلاغية المكونة من اسم + تعبيرة وصفية من جار ومجرور فقد كان عددها أقل نسبيًا؛ حيث لم يتجاوز إجمالي عدد صفحاتها خمس صفحات.

أما برمجية الاستعلام بالكلمة التي استُخدِمَت في ذخيرة لوب وذخيرة أخرى حجمها مائة وخمسة عشر ألف كلمة في مجال المراسلات الشخصية

فقد بَيّنت أن عدد مرات ظهور التركيبات الوصفية باستخدام الصيغة اسـم + تعبيرة وصفية من جار ومجرور أكثر بكثير من التركيبات الوصفية باستخدام الصيغة اسم + تعبيرة وصفية من الصلة والموصول؛ حيث وصلت نسبة تواجد الصيغتين في كل ألف كلمة على التوالي ٢٣,٣% مقابل ٥,٥%، بالإضافة إلى أن هناك دراسات قد أثبتت أن تركيب الاسم + التعبيرة الوصفية من جار ومجرور تعتبر من التركيبات الصعبة التي تواجه الدارسين الأجانب للغة الإنجليزية.

وهذا يشير إلى أن اهتمام القائمين على تعليم القواعد بظاهرة لغوية من عدمه يمر بأزمة حقيقية؛ حيث لا يتطابق تعليم القواعد مع الحقائق اللغوية الواقعية. وعند تأليف هذا النوع من المواد الدراسية لا ينبغي الاكتفاء بمراعاة درجة صعوبة الظواهر اللغوية وطرق تدريسها، بل ينبغي التفكير أيضًا في ظروف الاستخدام الواقعي لهذه الظواهر في اللغة. على هذا الأساس فقد تم تأليف كتاب (كولينز كوبويلد لقواعد اللغة الإنجليزية) على أساس التوافق مع قدر كبير من الحقائق اللغوية الموجودة في ذخيرة كوبويلد. وقد استغل هذا الكتاب تقنيات الاستعلام داخل الذخيرة لتقديم قدر كبير من التراكيب النحوية وأمثلة واقعية حولها.

الفصل الثاني

توظيف الإحصاء في علم الذخائر اللغوية

تعتبر الذخائر اللغوية مصدرًا مهمًّا لإجراء التحليل الكمي للغة. إلا أن استخدام الإحصاء الكمي في علم الذخائر اللغوية ليس بالأمر الذي يمكن إجراؤه بسهولة داخل المادة اللغوية المُمَثِّلة لمتن الذخيرة. والتقنيات الإحصائية المستخدمة في هذه الحالة لا تقتصر على إجراء التحليل الرياضي للبيانات اللغوية المعقدة واستخراج القواعد المنظمة لتلك البيانات اللغوية فحسب، بل يمكن استخدامها أيضًا في تفسير العلاقة بين كل من أسلوب الكتابة والتركيب اللغوي.

ويقدم هذا الفصل أكثر الطرق الإحصائية قيمة واستخدامًا في علم الذخائر اللغوية [2]. وفي هذا الجزء نُقَدِّم تعريفًا مختصرًا لأهم الطرق وأكثرها شيوعًا في علم الذخائر اللغوية؛ ونظرًا إلى عدم إمكانية حصر تفاصيل كل المنهجيات الإحصائية في هذا المقام؛ فسوف نكتفي بالحديث عن وسائل تطبيق تلك التقنيات الإحصائية في علم الذخائر اللغوية بشكل مبسط؛ على سبيل المثال، كيفية الاستخدام، والمغزى الحقيقي من استخدامها، مع عدم التعرض للتفاصيل الدقيقة لتطبيقها.

أولاً: إحصاء معدل التكرار

يُعتبر التعداد من أبسط الأعمال الإحصائية التي يتم تنفيذها على الذخائر اللغوية، بمعنى إجراء عملية عَدّ لعدد مرات ظهور بعض الظواهر اللغوية في أنواع معينة من النصوص. والتوصل إلى عدد مرات ورود المداخل اللغوية داخل الذخيرة، يساوي الحصول على عدد مرات تكرار تلك المداخل في الذخيرة بأكملها. ففي اللغة الإنجليزية، قد تكون تلك المداخل كلمة، أو تركيبًا لغويًا بعد حذف الإضافات الملحقة به، أو نوعًا معينًا من الكلمات، أما في اللغة الصينية فقد تكون تلك المداخل هي الرمز الصيني أو الكلمة الصينية. وسوف تكون الإشارة في هذا الفصل إلى الكلمة ما لم يكن هناك داعٍ للإشارة إلى نوع آخر من المداخل اللغوية. تكون خطوات التعداد كما يلي: معاينة كل كلمة من الكلمات داخل الذخيرة بصورة متوالية، وإذا كانت الكلمة قد ظهرت من قبل، يتم إضافة الرقم ١ على عدد مرات ظهور هذه الكلمة، وإلا يتم إضافة هذه الكلمة إلى قائمة الكلمات المكررة، ووضع الرقم ١ أمامها.

يُستخدم الرقم الذي تم التوصل إليه من خلال التعداد للإشارة إلى معدل تكرار الكلمات داخل الذخير بأكملها، وهذا الرقم له استخدامات كثيرة في علم الذخائر اللغوية. فإذا تمت فهرسة قائمة الكلمات المستخرجة من الذخيرة وفقًا لعدد مرات التكرار، يمكننا عمل دراسة عن ظروف توزيع المفردات داخل النصوص. على سبيل المثال، في نص تقني، إذا نقص استخدام المصطلحات التقنية في منطقة معينة من النص، وزاد استخدام هذه المصطلحات فجأة فيما بعد، فإن هذه الملاحظة تشير بصورة خاصة إلى الحدود الفاصلة بين أجزاء

النص؛ بمعنى أن يكون ذلك إشارة إلى نهاية مقدمة النص، وقد يكون ذلك إشارة إلى مقدمة قد كتبها أحد الأشخاص غير المتخصصين. أما ترتيب قائمة الكلمات حسب الترتيب الأبجدي فيُستَخدم بصورة أساسية في عمل قائمة مفهرسة لمفردات الذخيرة، وبذلك يمكن زيادة سرعة البحث عن الكلمات داخل الذخيرة؛ أما الترتيب التنازلي لقائمة المفردات حسب معدل التكرار، فيساعد في الدراسات القائمة على علاقة الأسلوب بالكلمات. وفيما يتعلق بمقارنة قائمة الكلمات ومعدل تكرارها على مستوى نوع معين من النصوص بمثيله على مستوى ذخيرة كبيرة، والتوصل إلى استخراج قائمة بالكلمات الأكثر تكرارًا في الذخيرتين، فيشير إلى أن تلك القائمة من الكلمات تتمتع بقدر من الثبات في الاستخدام. ويساعدنا في استبعاد تلك الكلمات ذات معدل التكرار الأعلى، وبالتالي تخمين الكلمات المفتاحية لنوع معين من النصوص(١).

ثانيًا: النسبة

النسبة التي تحتلها كلمة معينة داخل الذخيرة يُقصَد بها عدد مرات ظهور الكلمة داخل الذخيرة مقسومة على العدد الإجمالي لمجموع تكرارات

---

(١) على سبيل المثال إذا كان لدينا نص متخصص في مجال البيئة، وحصلنا على جدول تكراري تنازلي عن كلمات هذا النص، وعقدنا مقارنة بين مجموعة الكلمات الأعلى من حيث عدد مرات التكرار والكلمات المقابلة في ذخيرة أخرى عامة، وقمنا باستبعاد تلك الكلمات الأعلى من حيث التكرار من الجدول التكراري للنص المتخصص، فإن ما يتبقى من كلمات يُمثل بداية للتعرف على الكلمات المفتاحية التي تُستَخدَم في مجال البيئة. (المترجم)

كلمات الذخيرة بالكامل. وعلى الرغم من أن تعداد مرات التكرار يُعتبر طريقة فعالة للمعالجة الكمية للبيانات داخل الذخيرة، حيث إنها من الطرق المستخدمة دائمًا في الدراسات القائمة على ذخائر لغوية، فإن هذه الطريقة تشوبها بعض أوجه القصور. على سبيل المثال، تظهر أوجه قصور هذه الطريقة عند مقارنة مجموعتين من البيانات. فإذا أردنا الآن أن نقارن بين ذخيرتين إحداهما شفهية والأخرى تحريرية في اللغة الإنجليزية، فإن قائمة المفردات الخاصة بهاتين الذخيرتين ومعدل تكرار الكلمات بكل منهما يسجل عدد مرات ظهور كل كلمة في المجال اللغوي الطبيعي الذي خرج منه الإحصاء. وعندما يكون نطاق الذخيرتين غير متطابق من حيث الحجم، فمن الصعب الوثوق بجدول التكرار هذا لعمل المقارنة. وعلى الرغم من أن عدد مرات ظهور كلمة من الكلمات داخل إحدى الذخائر قد يكون أكبر من عدد المرات في الذخيرة الأخرى، فمن المحتمل أن تكون نسبة وجود هذه الكلمة في الذخيرة الأولى أقل من نسبة وجودها في الذخيرة الثانية. لو فرضنا أننا نقارن بين ذخيرتين للغة الإنجليزية أحدهما شفهية وتتكون من خمسين ألف كلمة، والثانية تحريرية وتتكون من خمسمائة ألف كلمة، وكان عدد مرات تكرر كلمة "boot" في الذخيرتين هو ٥٠، ٥٠٠ مرة على التوالي. فلو نظرنا من ناحية عدد مرات التكرار نجد أن كلمة "boot" قد تكررت في الذخيرة التحريرية أكثر من عدد مرات تكرارها في الذخيرة الشفهية، إلا أن الحقيقة ليست كذلك. والآن دعونا نحسب نسبة ظهور كلمة "boot" في الذخيرتين على السواء:

اللغة الشفهية: ٥٠/٥٠٠٠٠= ٠,١%

اللغة التحريرية: ٥٠٠/٥٠٠٠٠٠= ٠,١%

ومن الواضح أن كلمة "boot" لا يزيد معدل تكرارها في اللغة التحريرية بمقدار عشرة أضعاف كما يشير عدد مرات التكرار قبل حساب النسبة، بل إن معدل تكرار الكلمتين في الذخيرتين متساوٍ كما تشير معادلة النسبة التي تم حسابها. لذلك، فإنه عند مقارنة البيانات في ذخيرتين مختلفتين في الحجم، لا يمكن اللجوء إلى المقارنة السطحية لعدد مرات الظهور داخل الذخيرة، بل ينبغي حساب نسبة هذه الأرقام إلى رقم آخر؛ حيث يمكن أن يكون للرقم الناتج دلالة يمكن الاعتماد عليها. وفي ذلك الوقت يكون أسلوب الحساب المناسب هو:

$$\text{النسبة} = \frac{\text{عدد مرات ظهور الكلمة داخل الذخيرة}}{\text{إجمالي عدد كلمات الذخيرة}} \times ١٠٠$$

وهذه النسبة يتم التعبير عنها بالنسبة المئوية.

ثالثًا: أسلوب الاختبارات الإحصائية

على فرض أننا نريد مقارنة الإصدارين اللاتينيين لكل من إنجيل متا وإنجيل يوحنا، سنجد أن مقارنة حالات الاستخدام لصيغة المضارع "dicit" وصيغة الماضي "dixit" لفعل القول "to say" في الذخيرتين، علينا أن نبدأ بإحصاء عدد مرات تكرار كل كلمة في كل إصدار. وتكون نتيجة الإحصاء كما يلي:

	dicit	dixit
إصدار متا	46	107
إصدار يوحنا	118	119

ويمكننا أن نلاحظ من خلال الأرقام السابقة أن عدد مرات استخدام صيغة المضارع (dicit) في إنجيل يوحنا أكثر من عدد مرات استخدامها في إنجيل متا. وعلى فرض أن هناك عدم اتفاق بين الإنجيلين يتمثل في هاتين النقطتين، فإننا نحتاج إلى إثبات أن هذه الملاحظة ليست وليدة الصدفة من خلال العمليات الإحصائية. فنحن لا يمكننا الاكتفاء بالبيانات الموضحة بالجدول السابق للحصول على هذه النتيجة؛ ولكن الأمر يستدعي مستوى أعلى من التجريب، بمعنى إجراء اختبار إحصائي من أجل تقرير أوجه الاختلاف بين استخدام فعل القول (to say) في الإصدارين، وإقرار درجة احتمال الصدفة في وجود هذا الاختلاف.

يمكن لعلم الذخائر اللغوية اللجوء إلى أكثر من اختبار إحصائي للتوصل إلى ذلك، وهذه الاختبارات تشمل اختبار كا^2 (χ^2)، واختبار تي t وغيرها من الاختبارات. وحتى نوضح أهمية استخدام هذه الاختبارات في عمليات التحليل اللغوي، سنقدم شرحًا مبسطًا لاختبار كا^2 (χ^2)؛ وذلك نظرًا إلى أن اختبار كا^2 (χ^2) من أكثر الاختبارات الإحصائية استخدامًا؛ حيث يتمتع بالمميزات التالية:

(١) زيادة حساسيته تجاه البيانات مقارنة باختبار تي t.

(٢) لا يحتاج هذا الاختبار إلى توافر فرض "التوزيع الطبيعي" للبيانات، الذي لا يمكن الحصول عليه لبعض البيانات اللغوية.

(٣) سهولة حساب اختبار كا^2 (χ^2). ومن عيوب اختبار كا^2 (χ^2) أنه عندما تكون المادة التي يتم البحث عنها قليلة العدد فإن النتائج تفتقد إلى الموثوقية.

يُستَخْدَم اختبار كا^2 (χ^2) غالبًا في مقارنة الفرق بين عدد مرات التكرار المشاهَدَة داخل الذخيرة مع عدد مرات التكرار المُتَوَقَّعَة. وكلما اقترب عدد مرات التكرار المُشاهَدة مع عدد مرات التكرار المُتَوَقَّعَة كانت الظاهرة اللغوية تَحْدُث على سبيل المصادفة. وعلى العكس من ذلك، كلما زاد الفرق بين عدد مرات التكرار المُتَوَقَّعَة وعدد مرات التكرار المُشاهَدة، عَبَّرَ ذلك عن أن عدد مرات التكرار المُشاهَدة قد حدث نتيجة تأثير عوامل معينة وليس على سبيل الصدفة. وإذا تحدثنا عن المثال السابق، نجد أن هناك اختلافًا حقيقيًا بين إصداري الكتاب المقدس في طرق استخدام فعل القول (**tosay**).

بغض النظر عن التفاصيل الخاصة بخطوات حساب قيمة كا^2 (χ^2)[١]، وعلى فرض أننا انتهينا بالفعل من حساب قيمة كا^2 (χ^2) لصيغتي المضارع والماضي من فعل القول "**dicit**"، و"**dixit**"، ثم أردنا التعرف على درجة أهمية هذه القيمة في جدول قيم كا^2 (χ^2)؛ فقبل ذلك ينبغي تحديد قيمة درجة الحُرِّيَّة، التي تُحسَب من المعادلة التالية:

---

(١) لمزيد من التفاصيل عن اختبار كا^2 (χ^2) انظر مراجع الإحصاء الخاصة بالاختبارات اللامَعْلَمِيَّة.(المترجم)

درجة الحُريَّة = (عدد أعمدة الجدول التكراري – ١) * (عدد صفوف الجدول التكراري–١) وبعد ذلك يتم البحث في جدول قيم كا^2 (χ^2) عن قيمة الاحتمال المناظرة لتقاطع كل من درجة الحرية مع قيمة كا^2 (χ^2). وإذا اقتربت هذه القيمة من الصفر فإن هذا يعني أن الظاهرة حدثت بشكل عمدي ولم تحدث على سبيل المصادفة. وعلى العكس من ذلك إذا كانت القيمة أكبر من الصفر فهذا يعني أن الظاهرة حدثت على سبيل المصادفة. ونظرًا إلى أن درجة الاحتمال تقع بين ٠ و ١، فقد اتُفِقَ على اعتبار قيمة تقديرية تُعَبِّر عن أن النتيجة ذات مغزى إحصائي، وهذه القيمة المُتَّفَق عليها تقدر بـ ٠,٠٥ فإذا كانت قيمة الاحتمال التي يتم الحصول عليها من جدول الكشف أقل من ٠,٠٥ فإن هذا يدل على أن نسبة ٩٥% من المشاهدات تحدث عمدًا؛ وإلا فإن درجة موثوقية القيم الإحصائية التي يتم التوصل إليها لا تصل إلى نسبة ٩٥%.

والآن دعونا نحكم على الفرق بين الكلمتين في المثال السابق، وهل القيم الإحصائية التي تم الحصول عليها ذات مغزى أم لا؟ بالاستعانة بالجدول التكراري نجد أن قيمة كا^2 (χ^2) تساوي ١٤,٨٤٣، وبما أن الجدول التكراري السابق يحتوي عمودين وصفَّيْن، إذن درجة الحُرِّيَّة (٢–١) * (٢–١) = ١. وعند الكشف عن قيمة الاحتمال المقابلة لتلك القيمة في جدول توزيع كا^2 (χ^2) نجد أنها تساوي ٠,٠٠٠١ وهي بالطبع قيمة أقل من القيمة التقديرية ٠,٠٥ ولذلك يمكننا الحكم أن هذا الفارق يعبر في حقيقة الأمر عن أن هناك اختلافًا في استخدام فعل القول في إصداري الكتاب المقدس، وأن هذا الاختلاف لم يأت على سبيل المصادفة.

رابعًا: التصاحب اللغوي

التصاحب (collocation) من المفاهيم المهمة التي لها تطبيقات على نطاق واسع في علم اللغة. وباختصار، فإن التصاحب ما هو إلا نمط تظهر من خلاله المفردات بصورة اقترانية ويكون لذلك النمط خصائص محددة. ويرى كِجلمَر[51] (Kjellmer:1991) أن قاموس المفردات في عقل الإنسان لا يتكون فقط من كلمات مفردة، بل يحتوي العديد من الوحدات اللغوية التي تزيد عن الكلمة وترتبط فيما بينها بعلاقات نحوية. وهذه الوحدات منها ما هو ثابت، ومنها ما هو متغير. إن عملية تمييز أنماط التصاحب اللغوي داخل النصوص اللغوية (بالإضافة إلى ما قال به كِجْمِلَرْ من أنه نمط من أنماط التراكيب النحوية، فهو يوجد في تلك الكلمات التي دائمًا ما تظهر بمصاحبة كلمات معينة) تلعب دورًا في غاية الأهمية في مجال التأليف المعجمي؛ حيث يساعد ذلك في وضع معاني الكلمات وتحديد مجالات الاستخدام لكل منها وسياقاتها. وهذه المعلومات تلعب دورًا مماثلا من حيث الأهمية في علم معالجة اللغات الطبيعية وتعليم اللغات.

إن الكشف عن التصاحبات اللغوية لكلمة معينة داخل ذخيرة لغوية، إما أن يتم باستخدام المنهجيات الإحصائية، أو باستخدام المنهجيات القائمة على نظرية المعلومات.

١- كمية المعلومات المتبادلة واختبار زِد Z-score

إذا كان لدينا ذخيرة لغوية، يمكننا أن نكتشف أي الكلمات بينها درجة من درجات التصاحب الواضحة وذلك وفقًا لطبيعة البيانات التجريبية التي

يمكن استخراجها من تلك الذخيرة، ومن المؤكد وجـــود قـــوة اقترانيـــة ذات مغزى بين الكلمات المكونة لتلك التصاحبات، ولا يُعْتَبَر التجاور بينها تجاورًا على سبيل المصادفة. إن كمية المعلومات المتبادلـــة واختبـــار زِد Z-score دائمًا ما يتم اللجوء إليهما في الحكم عما إذا كانت هناك علاقة تصاحب بـــين كلمات معينة أم لا.

إن كمية المعلومات المتبادلـــة(١) (mutual information)[52] هـــي إحدى المفاهيم المعروفة في نظرية المعلومات؛ حيث يتم التعامل مع الكلمتين w_1، w_2 اللتين تُكَوِّنان تصاحبًا لغويًّا (ومن الممكن أيضًا أن يكون ذلك مـــع أي عنصرين تمت مشاهدتهما على أنهما يتلازمان معًا في الظهـــور) علـــى أنهما حدثان عشوائيان يحدثان معًا، ولحساب كمية المعلومات المتبادلة بـــين حدثين، ينبغي البدء بحساب احتمال تَحَقُّق هذين الحدثين معًـــا مـــن خـــلال المعادلة $p(w_1,w_2)$، بالإضافة إلى حساب احتمال ظهور كل كلمة بمفردهـــا $p(w_1)$، و $p(w_2)$، ويتم حساب كمية المعلومات المتبادلة من خلال المعادلة التالية:

$$M(w_1,w_2) = \log_2 \frac{p(w_1,w_2)}{p(w_1)p(w_2)}$$

إن المعنى الحقيقي لمصطلح كمية المعلومات المتبادلة هو مقدار حجم المعلومات التي يقدمها ظهور كلمة معينـــة لكلمـــة أخـــرى عنـــدما تظهـــر

---

(١) كمية المعلومات المتبادلة بين الحدث العشوائي w_1، والحدث العشوائي w_2 تُعَـــرَّف أنها مقدار الغموض الذي يتناقص حول w_1 بعد حدوث w_2. (المترجم)

بصحبتها. على سبيل المثال الكلمتان المتصاحبتان (riding, boot) تُكَوِّنَان وحدة لغوية مُجَمَّعَة، أما formula و borrowed فعلى الرغم من أنهما قد ظهرا معًا، كما في الجملة (... It is a formula borrowed from)، فإن هذا التصاحب قد حدث على سبيل المصادفة، ولا يوجد بين هاتين الكلمتين ارتباط من نوع خاص. وبصورة عامة، فكلما زادت قوة الارتباط بين كلمتين، زادت قيمة المعلومات المتبادلة بينهما؛ وإذا كان هناك ارتباط سلبي بين كلمتين (بمعنى أن ظهور إحداهما يمنع ظهور الثانية، والعكس صحيح)، فستكون قيمة المعلومات المتبادلة بينهما بالسالب. أما إذا كانت الكلمتان تظهران بصورة مستقلة (بمعنى عدم وجود علاقة بينهما)، ففي تلك الحالة ستكون قيمة المعلومات المتبادلة تساوي صفرًا. وبعبارة أخرى، فإن الكلمتين اللتين بينهما قيمة معلومات متبادلة كبيرة يزداد احتمال وجود اقتران ذي معنى بينهما، أما عندما تقترب قيمة المعلومات المتبادلة بين كلمتين من الصفر أو تقل عنه، فلا يمكن أن يكون هناك تصاحب بينهما.

أما القيم التي يقدمها اختبار زِد (Z-score) فتشبه تلك التي تقدمها معادلة كمية المعلومات المتبادلة. فبالنسبة إلى كلمة معينة داخل نص، يُقَدِّم اختبار زِد (Z-score) مقارنة بين المشاهدات الحقيقية والمشاهدات المُتَوَقَّعَة للكلمات الأخرى التي تظهر في السياق المصاحب لتلك الكلمة. وكلما زادت قيمة اختبار زِد (Z-score) لكلمة من الكلمات مع كلمة أخرى يتم تحديدها، زادت القوة التصاحبية بينهما (أو القوة الاقترانية بينهما)، بمعنى زيادة إمكانية وجود اقتران له دلالة بين هاتين الكلمتين. ولا يُستَخدم اختبار زِد (Z-score)

كثيرًا في علم الذخائر اللغوية، إلا أن هناك برمجية تسمى **TACT** للفهرسة السياقية استعانت بتلك المنهجية.

٢- تطبيقات كمية المعلومات المتبادلة واختبار زِد (Z-score)

الاستخدام الرئيس لهما هو استخراج الوحدات اللغوية المُركَّبَة من أكثر من كلمة، ولا يشمل ذلك التعبيرات اللغوية الشائعة فقط مثل "**cock and bull**" بل يمتد ليشمل التعبيرات الاسمية أيضًا، مثل "**temporal mandibular joint**" فهذا الأخير يُعتبر تعبيرًا اصطلاحيًا، وبجانب استخدام نتائج هذين الاختبارين في صناعة المعاجم، فإن لهما استخدامًا مهمًّا في مجال الترجمة؛ حيث يمكن من خلالهما بناء قاعدة معارف تفصيلية عن المصطلحات المستخدَمة في مجال معين.

أما ثاني أهم التطبيقات التي تُستَخدم فيها كمية المعلومات المتبادلة واختبار زِد (**Z-score**) فيتمثل في إمكانية المساعدة في إزالة اللبس اللغوي (**Word Sense Disambiguation**) والذي يُطلَّق عليه اختصارًا **WSD**). وهذا يختلف عن التطبيق السابق، لأننا في تلك الحالة نكون بصدد استخراج مجموعة التصاحبات اللغوية المألوفة لكلمة من الكلمات، فإذا أردنا أن نُحَدِّد أهم التصاحبات التي تأتي مع كلمة ما، فمن الممكن أن يكون ذلك في إطار الاحتمالات التالية:

(١) نجمع التصاحبات التي تشترك في كلمة واحدة في مجموعات، الأمر الذي يساعد علماء اللغة في التعرفِ الآلي على المعاني المختلفة لتلك الكلمة من خلال قائمة مفهرسة للتصاحبات. فكلمة "**bank**" على سبيل المثال

يمكن أن تُكَوِّن تصاحبًا لغويًا مع كلمات معينة في مجال الجغرافيا (مثل كلمة river) ولكنها تُكَوِّن تصاحبًا آخر مع كلمات مختلفة في مجال البنوك والمال (مثل كلمة investment)، ومن هنا يمكننا التمييز بين دلالتين مختلفتين لكلمة "bank"، وفي الوقت ذاته يتم إجراء مقارنة بين كلمتين مختلفتين من خلال علاقاتهما الاقترانية بكلمات أخرى للتعرف على مدى الاختلاف بين هاتين الكلمتين من حيث الاستخدام. وفي هذا الإطار، أجرى العالم ليتش (Leech:1992)[53] تجربة لمقارنة الاختلاف في الاستخدام بين الكلمتين "strong"، و"powerful" في اللغة الإنجليزية. فاستخدم نظرية المعلومات المتبادلة لاستخراج الاقترانات المصاحبة لكل من هاتين الكلمتين داخل الذخيرة اللغوية، وكانت النتيجة أنه اكتشف اختلاف هاتين الكلمتين من حيث القدرة الاقترانية. فكلمة "strong" تقترن مع كل من "northerly"، و"showings"، و"behaviour"، و"currents"، و"supporter"، وما إلى ذلك، أما كلمة "powerful" فتقترن مع كلمات مثل "tool"، و"minority"، و"neighbor"، و"symbol"، و"figure"، "weapon"، وما إلى ذلك. وعلى الرغم من أن هناك بعض التصاحبات التي تم حصرها فإنه لا يمكن أن يُطلَق عليها اقتران بالمعنى الدقيق للكلمة؛ ولكن يمكن أن يتضح منها وجود اختلافات جوهرية بين هاتين الصفتين من حيث الاستخدام.

وهناك استخدام آخر مهم لنظرية المعلومات المتبادلة، ألا وهو المساعدة في دراسة العلاقة الدلالية بين ذخيرتين لغويتين متوازيتين ثنائيتي اللغة على مستوى الأسطر. فعلى فرض أن لدينا ذخيرتين ثنائيتي اللغة وتم عمل توازٍ بينهما على مستوى الأسطر، بالإضافة إلى عمل توازٍ على

مستوى الجمل؛ بمعنى أننا يمكننا اختيار جملة ما داخل إحدى الذخيرتين، وتحديد الجملة المترجمة عنها في الذخيرة الموازية لها. فبناءً على ذلك، يمكننا عن طريق الحساب أن نتوصل إلى الكلمات وترجمتها داخل كل جملة من جمل الذخيرة.

خامسًا: النماذج اللغوية

١- نموذج العنصر N (نموذج الرُّتْبة N-1 لماركوف)

على فرض أن لدينا سلسلة من العلامات S تتكون من عدد L من العلامات، وهذه العلامات على التوالي يرمز لها بالرموز $w_1, w_2, \ldots, w_L$، فإن النموذج اللغوي الذي يَعتَمِد على تواتر تلك العلامات معًا يرى أن احتمال ظهور عناصر السلسلة S معًا يُحسَب من المعادلة التالية:

$$P(S) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_2 w_1)\ldots P(w_L \mid w_1 \ldots w_{L-1}) = \prod_{i=1}^{L} P(w_i \mid w_1 \ldots w_{i-1})$$

وعند بناء فرض مستقل وفقًا للمعادلة السابقة، بمعنى افتراض أن ظهور كل كلمة w_i في السلسلة الكلامية S يرتبط فقط بالكلمة السابقة لها؛ التي تحتل الموقع $N-1$ وذلك وفقًا للصيغة $w_{i-(N-1)} \cdots w_{i-1}$، بالإضافة إلى عدم وجود علاقة بينها وبين الكلمات الأخرى خارج نطاق $N-1$. فيمكن التعبير عن المعادلة السابقة كما يلي:

$$P(S) \approx \prod_{i=1}^{L} P(w_i \mid w_1 \ldots w_{i-1})$$

$$= P(w_1)P(w_2 \mid w_1)...P(w_{N-1} \mid w_1...w_1...w_{N-2}).\prod_{i=N+1}^{L} P(w_i \mid w_l...w_{N-1})$$

إن النماذج النحْوية المكَوَّنة من عدد N من الوحدات تـساوي نمـوذج الانتقال بخطوة لماركوف N-1.

النماذج النحوية الأخرى الأكثر شيوعًا؛ التي تنتمي إلى نموذج الانتقال بعدد N من الخطوات هي: نموذج الانتقال بخطوتين ونموذج الانتقال بـثلاث خطوات وذلك عندما N=2، وN=3 على التوالي. بمعنى أن احتمال ظهور أي كلمة ينحصر فقط في ظهورها مع كلمة أو كلمتين سابقتين لها داخل الـنص. وعلى ذلك فإن معادلة حساب تلك الاحتمالات يُعَبَّر عنها بالصيغ التالية:

معادلة حساب النموذج النحوي الثنائي:

$$P(S) = P(w_1)\prod_{i=2}^{L} P(w_i \mid w_{i-1})$$

معادلة حساب النموذج النحوي الثلاثي:

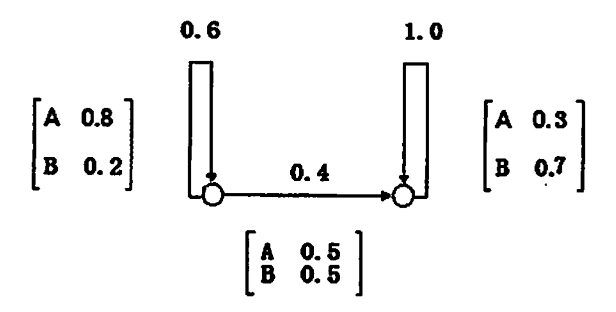
$$P(S) = P(w_1)P(w_2 \mid w_1)\prod_{i=2}^{L} P(w_i \mid w_{i-2}w_{i-1})$$

٢- نموذج ماركوف الكامن

Hidden Markov Model (HMM)

نموذج ماركوف الكامن HMM هو ائتلاف مجموعة مـن الحـالات تربطها سلسلة من التحولات التي تَكُون دائمًا في إطار احتمالين: الأول هـو

احتمال التحول (transition probability) وهو الذي يُعطي احتمال حـــدوث هذا التحول، والثاني المتوالية المُعَبِّــرَة عــن كثافــة مخرجــات الاحتمـــال (O Output Probability Density Function) (PDF) وهو الذي يُعَرِّف احتمال خروج كل رمز من قائمة حروف أبجدية محدودة العدد كما يظهر لنا من الشكل (٣-٢).



شكل (٣-٢): نموذج ماركوف كامن مُبَسَّط

(حالتان، ورمزان مُخْرَجان A، وB)

التعريف الشكلي لنموذج ماركوف الكامن كما يلي:

(أ) اقتران الحالة $\{S\}$: ويشمل حالة بداية S_i، وحالة نهاية S_F؛

(ب) اقتران التحول a_{ij}: $\{a_{ij}\}$: حيث تُعَبِّر a_{ij} عن احتمال التحول مــن الحالة i إلى الحالة j.

$$a_{ij} = P(X_{t+1} = j \mid X_t = i), \forall_i, j, a_{ij} \geq 0, \sum_j a_{ij} = 1$$

(ج) التنظيم الاحتمالي للمخرجات b_{ij}: $\{b_{ij}(k)\}$: حيث تُعَبِّــر b_{ij} عــن احتمال ظهور العلامة k في أثناء التحول من الحالة i إلى الحالة j.

$$b_{ij} = P(Y_t = k \mid X_t = i, X_{t+1=j)}, \forall_i, j, k, b_{ij} \geq 0, \sum_k b_{ij}$$

حيث تُعَبِّر $X_t = j$ في المعادلة السابقة عن أن الحالة j تحدث فــي اللحظة t، و $Y_t = k$ تُعَبِّر عن ظهور العلامة k في اللحظة t.

وإذا كان لدينا نموذج ماركوف كامن **HMM** يُطلَــق عليــه **M** فــإن احتمال أن يُولِّد هذا النموذج سلسلة من العلامات يُطلَق عليها y_1^t يُحسَب من خلال المعادلة التالية:

$$P(Y_1^y) = \sum_{x_1^{T+1}} P(X_1^{T+1} = x_1^{T+1}) P(Y_1^T = y_1^T \mid X_1^{T+1} = x_1^{T+1})$$

والمعنى المباشر لتلك المعادلة هو: حصر مسارات التحول في سلسلة العلامات y_1^T؛ التي تنشأ بأطوال مختلفة من T ويُرْمَــز لهــا بــالرمز x_1^{T+1} (ويُطلَق عليها أيضًا اسم سلسلة ماركوف)، بالإضافة إلى البحث عن احتمال حدوثها، ومن بين ذلك فإن احتمال حدوث كل مسار من مــسارات السلــسلة x_1^{T+1} يتم حسابه من خلال الجمع بين احتمال التغير في هذا المسار بالإضافة إلى احتمال المخرجات. إن سلسلة ماركوف X وتَسَلْسُل العلامات Y الخارجة منها جميعها تنشأ من نموذج واحد كامن لماركوف **HMM**؛ ولكــن تسلــسل المخرجات Y يُمْكِن مشاهدته مباشرة، أما تسلسل الحــالات X فهــو كــامن **(hidden)**.

يوجد فرضان ضمن مستوى واحد مــن نمــاذج مــاركوف الكامنــة **HMM**؛ وهما:

فرض ماركوف (Markov assumption)

$$P(X_{t+1}=x_{t+1} \mid X_1^t=x_i^t)=P(X_{t+1}=x_{t+1} \mid X_t=x_t)$$

حيث تُعَبِّر X_i^j في المعادلة السابقة عن تسلسل الحالات، التي يرمز له بالرموز $X_i, X_{i+1}, \ldots, X_j$. وفرض ماركوف يُعَبِّر عن أن احتمال وصول سلسلة ماركوف إلى حالة محددة لا يحدث إلا في اللحظة t المرتبطة بحالة سلسلة ماركوف.

الفرض المستقل عن المخرجات (output-independence assumption):

$$P(Y_t=y_t \mid Y_1^{t-1}=y_1^{t-1}, X_1^{t+1}=x_1^{t+1})=P(Y_t=y_t \mid X_t=x_t, X_{t+1}=x_{t+1})$$

حيث تُعَبِّر Y_i^j في المعادلة السابقة عن سلسلة المخرجات $Y_i, Y_{i+1} \ldots Y_j$.

إن الفرض المستقل عن المخرجات يُعَبِّر عن أن احتمال ظهور مجموعة علامات معينة في اللحظة t يرتبط ارتباطًا وثيقًا بمقدار التحول (من x_t إلى x_{t+1}) الذي يحدث في اللحظة نفسها.

وفي إحدى نماذج ماركوف الكامنة HMM، نجد أن احتمال توليد النموذج M من السلسلة y_1^T يُحْسَب من المعادلة:

$$P(Y_1^T=y_1^T)=\sum_{x_1^{T+1}} \prod_{t+1}^{T} P(X_{t+1}=x_{t+1} \mid X_t=x_t) P(Y_t= y_t \mid X_t=x_t, X_{t+1}=x_{t+1})$$

٣- النموذج النحوي للعنصر N، مقارنة بين كل من النموذج المُعتمِد على النحو، ونموذج ماركوف الكامن HMM

إن طريقة حساب النمط المُعتَمِد عل النحو للعنصر N (نموذج الرُتْبة N-1 لماركوف) بسيط وعملي ويُسْتَخْدَم على نطاق واسع في تمييز الأصوات اللغوية والأنماط اللغوية في مجال اللغات الطبيعية. إلا أن أهم صفة تتميز بها اللغات الطبيعية هي الاتجاه نحو التركيب، وعلى العكس من ذلك نجد أن النموذج النحوي للعنصر N ما هو إلا نموذجًا لغويًا يعتمد على إظهار العلاقة بين العلامات بصورة خطية، ولا يمكنه إلا تمييز العلامات التي تُعَبِّر عن معلومات البنية السطحية للغة (عادةً ما يكون ذلك من خلال العلامات المُعَبِّرَة عن الرموز والكلمات وأنواعها النحوية) وما بينها من مظاهر تُعَبِّر فقط عن علاقات التجاور في النصوص اللغوية؛ ومن ثَمَّ لا يمكنها استقراء حالات التراكيب اللغوية؛ ولذلك فإن النموذج النحوي للعنصر N يُسْتَخْدَم على نطاق محدود في التعبير عن اللغات الطبيعية بطبيعتها المركبة.

يُعتبر نموذج ماركوف الكامن HMM هو الشكل المُطوَّر من نموذج ماركوف التقليدي. فنموذج ماركوف التقليدي يكتفي بوصف التحُول الذي يحدث لكل حالة من الحالات اللغوية بشكل عشوائي، أما نموذج ماركوف الكامن HMM فيصف حدثين عشوائيين: الحدث العشوائي الأول يصف احتمال ظهور العلامة اللغوية والحالة التي تكون عليها؛ أي أن المُخْرَجات تكون دالة في الحالة؛ أما الحدث العشوائي الثاني فلا يصف إلا علاقة التحُول

التي تحدث بين الحالات. وبالنسبة إلى المستخدم العادي، فإن ما يــراه هــو المخرجات فقط، ولا يمكنه أن يشاهد التحول الذي يحــدث بــين الحــالات المختلفة؛ أي أن التحول بين الحالات يكون مخفيًّــا. إن نمــوذج مــاركوف الكامن HMM يشبه الطرق التقليدية في حساب الاحتمالات [٥٤]، من حيــث محدودية الحالات التي يمكن التعامل معها، ومن ثَم عــدم إمكانيــة وصــف مستويات التركيب المختلفة في اللغات الطبيعية[٥٥].

٤- تطبيقات نموذج ماركوف الكامن HMM في الذخائر اللغوية

يمكن التعرف على التطبيقات الحقيقية لنموذج ماركوف الكامن **HMM** في معالجة اللغات الطبيعية من خلال ما يتم في عمليات ترميز ذخيرة لغوية.

وصف المشكلة

نفترض أن لدينا سلسلة من الكلمات $w_1, w_2, ..., w_T$ ، ونريد توصيف تلك السلسلة من حيث الأنواع النحوية لها $c_1, c_2, ..., c_T$ ، ونظرًا إلى انتشار ظاهرة اللبس اللغوي في تمييز الأنواع النحوية للكلمات، فمــن الممكــن أن يقابــل السلسلة الواحدة من الكلمات عدة سلاسل من أنواع الكلمات، هذا بالإضــافة إلى أن سلسلة أنواع الكلمات التي نريد الحصول عليها ستجعل قيمة المعادلة **PROB** $(c_1, c_2, ..., c_T \mid w_1, w_2, ..., w_T)$ تعادل قيمة أكبر سلــسلة مــن الأنواع النحوية للكلمات[٥٦].

وباستخدام قانون بايز Bayes[1] للاحتمالات يمكننا كتابة المعادلة السابقة بالصيغة التالية:

$$\frac{PROB(w_1, w_2, \ldots, w_T \mid c_1, c_2, \ldots, c_T) \times PROB(c_1, c_2, \ldots, c_T)}{P(w_1, w_2, \ldots, w_T)}$$

حيث يُطلَق على $PROB(w_1, w_2, \ldots, w_T \mid c_1, c_2, \ldots, c_T)$ اسم معادلة المعلومات المتعلقة بالمفردات، ويُطلَق على $PROB(c_1, c_2, \ldots, c_T)$ النموذج اللغوي، ونظرًا إلى ثبات المقام بالنسبة إلى سلاسل الكلمات المتساوية فإن المعادلة السابقة يمكن اختصارها إلى المعادلة التالية التي تَحْسِب أكبر سلسلة من أنواع الكلمات.

$$PROB(w_1, w_2, \ldots, w_T \mid c_1, c_2, \ldots, c_T) \times PROB(c_1, c_2, \ldots, c_T)$$

ويمكننا وضع مستوى أعلى من الفروض للمعادلة السابقة: إن احتمال ورود الكلمة الحالية يتم التوصل إليه من النوع النحوي للكلمة، والنوع النحوي لهذه الكلمة مرتبط فقط بنوع الكلمة السابقة لها. وفي النهاية يتم التعبير عن الموضوع بالكامل من خلال المعادلة التالية:

$$T^* = \arg\max_{c_1, c_2, \ldots, c_T} p(w_1)\, p(w_1 \mid c_1)\, p(c_1) \prod_{i=2}^{T} p(c_i \mid c_1, c_2, \ldots, c_{i-1})\, p(w_i \mid c_i)$$

حيث تشير T^* إلى سلسلة الكلمات التي يتم ترميزها في النهاية، وتشير P(.) إلى الاحتمال.

---

(1) قانون بايز هو إحدى النتائج المهمة لنظرية الاحتمالات ويقوم بحساب التوزيع الاحتمالي الشرطي للمتغير العشوائي A بمعلومية المتغير العشوائي B. (المترجم)

من المعادلة السابقة يمكننا أن نستخرج المستوى الأول والثاني من نموذج ماركوف الكامن **HMM**؛ حيث إن المستوى الأول لنموذج **HMM** يُعَبِّر عن أن النوع النحوي للكلمة الحالية لا يرتبط إلا بنوع الكلمة السابقة عليها. وتكون المعادلة بالتفصيل كما يلي:

$$T^* = \arg\max_{c_1, c_2, \ldots, c_T} p(c_1)p(w_1 \mid c_1)\prod_{i=2}^{T} p(c_i \mid c_1, c_2 \ldots, c_{i-1})p(w_i \mid c_i)$$

حيث تُعَبِّر $p(c_1 \mid c_{i+1})$ عن احتمال تغير الحالة في نموذج HMM، وتُعَبِّر $P(w_i \mid c_i)$ عن احتمال توليد الكلمات.

وبذلك يكون التعبير عن مسألة ترميز أنواع الكلمات قد تم من خلال أفضل صيغة لنموذج ماركوف الكامن **HMM**، ويمكن التأكد من صحة النتائج السابقة فيما يتعلق بحساب احتمال التغير في الحالة واحتمال توليد الكلمات من خلال دراسة ذخيرة لغوية سبق ترميزها.

الفصل الثالث

برامج الفهرسة وتطبيقاتها

يقدم هذا الفصل تعريفًا لبرمجيتين إحصائيتين يتم استخدامهما في تطوير الذخائر اللغوية قام بتطويرهما العالم اللغوي سنكلير.(Sinclair) وتقوم هاتان البرمجيتان بعرض الكلمة التي يتم الاستعلام عنها مع السياقات التي وردت فيها الكلمة داخل الذخيرة في كل مرة وردت فيها. ويتمثل الاختلاف بينهما في أن كلا منهما تقدم المعلومات التي يتم الاستعلام عنها مُرَتَّبَة حسب معايير إحصائية تختلف عن الأخرى؛ وذلك لمساعدة المستخدم في تحليل الظواهر اللغوية. فبرمجية كولوكيت **Collocate** تقوم بحساب الكلمات التي تتوارد مع الكلمة موضع البحث بصورة مباشرة، أما برمجية تيبيكال **Typical** فتحسب التوارد مع الكلمة موضع البحث على مستوى السطر بأكمله بغض النظر عن تجاور الكلمتين معًا. وقد تجاوزت النتائج التي حققتها هاتان البرمجيتان كل التوقعات المبدئية، وتمتعت كل منهما بمكانة عالية من حيث شيوع الاستخدام في مجال تطوير الذخائر اللغوية. ويشير المرجع [٥٧] بالتفصيل إلى العمل الذي قام به سنكلير من خلال هاتين البرمجيتين، وسوف نشرح ذلك بالتفصيل في الأجزاء التالية:

أولاً: برمجية كولوكيت COLLOCATE

تقوم برمجية كولوكيت **Collocate** بدايةً بحساب عدد مرات تكرار كلمة ما داخل جدول الكلمات المُفَهْرَس الذي تم استخراجه، بالإضافة إلى

احتمال ظهور تلك الكلمة داخل االذخيرة اللغوية. وبناءً على هذه المعلومات تتمكن البرمجية من حساب درجة وضوح علاقة التصاحب بين هذه الكلمة وكلمة أخرى يتم تحديدها سابقًا. والمقصود بمدى الوضوح هنا النسبة بين احتمال ظهور الكلمة المُرَشَّحَة لتكوين التصاحب داخل الذخيرة بأكملها وبين احتمال ظهورها داخل النص موضع الدراسة.

١- طريقة الحساب

في البداية يتم حساب عدد مرات ظهور الكلمة المطلوبة داخل الذخيرة، ثم تُحْسَب القيم المشاهدة والقيم المتوقعة لكل كلمة من الكلمات المُكَوِّنة لكل سجل من سجلات الملف المُفَهْرَس. وعند حساب التكرار يمكن اختيار إحدى الطرق التالية: بعد إهمال حالة الأحرف الكبيرة والصغيرة في الكلمات الإنجليزية المُفْرَدة، يمكن مراقبة نوع الكلمات بعد حذف علامات التصريف، ويمكن أيضًا مراقبة الكلمات موضع الدراسة مع الكلمات التي تتصاحب معها سواء من جهة اليسار أو اليمين أو كلاهما معًا.

المدخلات التي يتم الحساب عليها: ملف التحليل الإحصائي السياقي **concordance file** عن سياقات الكلمة موضع البحث يَنْتُج عن البرمجية. فعندما تظهر الكلمة موضع البحث داخل الذخيرة، تقوم البرمجية بإنشاء قائمة من مجموعة من السجلات تحتوي الكلمات التي تصاحبت مع تلك الكلمة المفتاحية يمينًا ويسارًا وفقًا لامتداد معين من الكلمات يُطْلَق عليه نافذة. وجميع السجلات المُحتَواة في تلك القائمة يتكون منها ملف التحليل الإحصائي السياقي للكلمة **concordance file**.

في البداية يتم الحصول على قائمة بها معدل تكرار جميع الكلمات عن طريق إحصاء يتم على ذخيرة واسعة النطاق.

بعد ذلك يتم حساب درجة وضوح التصاحب التي يُشار إليها بـــالرمز S بين الكلمة موضع البحث وجميع الكلمات الأخرى ويشار إليها بالرمز W وتكون معادلة الحساب كما يلي:

S=	احتمال ظهور الكلمة w داخل ملف الفهرسة / احتمال ظهور الكلمة w داخل الذخيرة بأكملها	$= \frac{OF}{EF}$

حيـــــــــــث $OF = freq_{span}(w)/N_{span}$، و $EF = freq_{corpus}(w)/N_{corpus}$، و $freq_{span}(w)$، و $freq_{corpus}$ تُعَبِّر كلٌّ منها عن عدد مرات تكرار الكلمة w في ملف سياقات الكلمة وداخل الذخيرة بأكملها على التـــوالي، أمـــا N_{span}، و N_{corpus} فتُعَبِّر كلٌّ منها عن عدد الكلمات في ملف سياقات الكلمة والـــذخيرة بأكملها على التوالي.

المخرجات: عند عرض المخرجات يتمّ ترتيب الكلمـــات المـــصاحبة للكلمة التي يتم الاستعلام عنها وفقًا للقيمة المحدَّدَة لوضوح درجة التصاحب في المعادلة السابقة، ويحتوي كل سجل أربع خانات تضم المعلومات التالية:

أ) التصاحبات: حيث تُعرّض الكلمات التي يتوافق ظهورها مع الكلمة موضع الاستعلام.

ب) عدد مرات تكرار الكلمة: حيث يُعـرّض عـدد مـرات تكـرار التصاحب داخل الذخيرة.

ج) معدل التكرار المتوقع: حيث يُعرَض معدل التكـرار الاحتمـالي المتوقع للتصاحب في إطار سياق له طول محدد.

د) معدل التكرار الحقيقي: حيث يُعرّض عدد مرات الظهور الحقيقـي للكلمات المتصاحبة مع الكلمة موضع البحث داخل ملف السياقات.

وعند تنفيذ هذه العملية، يكون هناك اختياران:

(١) حساب أو إهمال حالة الأحرف.

(٢) حساب أو إهمال معلومات موضع الكلمة. وسوف يتم شرح تطبيق هذه العمليات من خلال الأمثلة الثلاثة التالية. على فرض أن الكلمة التي نريد الاستعلام عنها هي كلمة "arms"، وأن طول السياق هـو أربـع كلمـات، سيعرض الجدولان (٣-١)، و(٣-٢) النتائج كلٌ على حدة؛ مرة مع إغفـال التغيرات الصرفية، ومرة أخرى مع احتساب التغيـرات الـصرفية؛ حيـث يَعْرِض العمود الأول من كل جدول الكلمات التي ترد في حالة تصاحب مـع هذه الكلمة، ويَعْرِض العمود الثاني والرابع على التوالي عدد مرات ظهـور الكلمة المتصاحبة مع كلمة "arms" في كل من الذخيرة وملف السياقات على التوالي. أما العمود الثالث فيعرض قيمة وضوح درجة التصاحب التـي تـم

حسابها من خلال البرمجية. وبعد ذلك تكون النتيجة التـــي يقــدمها برنـــامج كولوكيت Collocate هي عرض النتائج مُجمَّعة في كل عمود من الأعمدة.

جدول (٣-١): مع إغفال التغيرات الصرفية

الكلمات المُتصاحبة مع كلمة "arms"	عدد مرات ظهور الكلمة المُتصاحبة في الذخيرة	قيمة وضوح درجة التصاحُب	عدد مرات ظهور الكلمة المُتصاحبة داخل ملف السياقات
Caches	75	0.073	30
Outstreched	372	0.364	97
Cache	254	0.248	66
Cradled	168	0.164	38
Flailing	237	0.232	37
Embargo	3427	3.352	527
Folded	1462	1.430	195
Ammunition	1910	1.868	154
Shipments	1238	1.211	96
Treaties	807	0.789	60
Legs	8507	8.320	594
Waving	1741	1.703	116
Aloft	409	0.400	23
Torso	406	0.397	21
Flung	1176	1.150	59
Elbows	613	0.600	30

جدول (٣-٢): مع اعتبار التحولات الصرفية

الكلمات المُتصَاحبة مع كلمة "arms"	عدد مرات ظهور الكلمة المُتصَاحِبة في الذخيرة	قيمة وضوح درجة التصَاحُب	عدد مرات ظهور الكلمة المُتَصَاحِبة داخل ملف السياقات
Buildup	**441**	**0.431**	**21**
Cache	**329**	**0.322**	**96**
outstretched	**372**	**0.364**	**97**
Flail	**3791**	**3.708**	**556**
Ammunition	**1910**	**1.868**	**154**
Aloft	**409**	**0.400**	**23**
Fold	**4707**	**4.604**	**236**
Cradle	**1085**	**1.061**	**53**
Buildup	**441**	**0.431**	**21**
Torso	**475**	**0.465**	**22**
Strategic	**5712**	**5.587**	**256**
Fling	**1857**	**1.816**	**83**
Reduction	**8191**	**8.011**	**343**
Conventional	**7157**	**7.000**	**292**
Gent	**615**	**0.601**	**25**
Smuggle	**2024**	**1.980**	**78**

من الجدول (٣-٢) يمكننا ملاحظة وضوح درجة التصاحب بين كلمة "cashe" وكلمة "arms" بدرجة كبيرة، فقد وصل عدد مرات ظهور "cache" داخل الذخيرة إلى ٣٢٩ مرة، ومن الجدول (٣-١) يمكننا أن نلاحظ أن هذا الرقم يمثل مجموع ظهور كل من كلمتي "cache"، و"caches" داخل الذخيرة. فالتصاحبات اللغوية بين الكلمات داخل الذخيرة في الجدول (٣-٢) قد تم حسابها بين الكلمات التي تظهر معًا بعد استبعاد الإضافات الصرفية لكل كلمة.

أما الجدول (٣-٣) فيعرض معلومات عن التصاحب بين الكلمات من حيث موضع الكلمة المتصاحبة من الكلمة الأساسية، وفي تلك الحالة احتوى الجدول عمودًا إضافيًّا هو العمود الخامس الذي يحتوي معلومات تشير إلى موضع الكلمة المصاحِبة.

جدول (٣-٣): يوضح اتجاه التصاحب (تصاحب أيسر أو تصاحب أيمن)

موضع الكلمة المُصَاحِبة	عدد مرات ظهور الكلمة المتصاحبة داخل ملف السياقات	قيمة وضوح درجة التصاحُب	عدد مرات ظهور الكلمة المتصاحبة في الذخيرة	الكلمات المتصاحِبة مع كلمة "arms"
No left	27	0.073	75	Caches
No right	36	0.164	168	Cradled
Left discarded	65	0.364	372	Outstretched
Left discarded	488	3.352	3427	Embargo
	66	0.248	254	Embargo
No left	153	1.868	1910	Ammunition
No left	57	0.789	807	Treaties
	195	1.430	1462	Folded
No left	82	1.211	1238	Shipments
No right	105	1.703	1741	Waving
No left	23	0.400	409	Aloft
Left discarded	468	8.320	8507	Legs
No right	249	5.587	5712	Strategic
No left	82	1.847	1888	Reductions
No right	122	2.817	2880	Lifting
No right	276	7.000	7157	Conventional
No left	241	6.165	6303	Reductions
No right	50	1.282	1311	Supplying
No right	44	1.150	1176	Flung
No left	33	0.911	913	Negotiator
No left	40	1.229	1257	Explosives
No left	22	0.721	7371	Shipment

تشير "no left" إلى أن التصاحب قد تم من جهة اليمين، أما "no right" فتشير إلى أن التصاحب قد تم من جهة اليسار. وإذا كان عدد مرات ظهـور الكلمة المتصاحبة في جهة من الجهات أكبر من ثلثـي قيمتـه فـي الجهـة الأخرى، يتم إهمال التصاحب على الجهة الأخرى التي تحقق عـدد مـرات تصاحب أقل. ويشير كل من "left discarded"، و"right discarded" إلـى جهة التصاحب التي تم إهمالها سواء كانت جهة اليسار أو جهة اليمين. أمـا إذا كان الفرق غير كبير بين عدد مرات التصاحب في الجهتين فتُترك الخانة المخصصة لذلك فارغة.

ثانيًا: برمجية TYPICAL

تعتمد برمجية تيبيكال Typical أساسًا على حساب درجـة وضـوح ظهور الكلمات معًا في سطر الفهرسة، ويُستخدم ذلك في تقدير درجة وضوح تكرار سطر الفهرسة بأكمله، ويساعد ذلك في استخراج أمثلـة واقعيـة ذات طبيعة خاصة من داخل الذخيرة. وتكون مدخلات البرمجية عبارة عن ملف توافق وقوع عن كلمة من الكلمات الموجودة في الـذخيرة، بالإضـافة إلـى جدول عن معدل تكرار كل كلمة من الكلمات داخل الذخيرة. وبعد ذلك، تقوم البرمجية بمعالجة كل سطر من أسطر الملف المفهرس وتَـستَخرِج أفـضل سطر يمثل الشكل القياسي للاستخدام. كان التصميم المبدئي لتلك البرمجيـة يهدف إلى محاولة البحث عن الأمثلة النموذجية التي تمثل الاستخدام الحقيقي للغة؛ وذلك كي تكون أداة مساعدة للمعجميين في صناعة المعـاجم؛ حيـث تُمكِّنهُم من البحث عن أمثلة لغوية حقيقية ذات موثوقية. إلا أن استخدام تلك

البرمجية على أرض الواقع أسفر عن استخدامات أكثر رحابة لم تكن متوقعة في مجال إزالة اللبس الدلالي بين الكلمات.

من أحد الفروض التي تم استخدامها في تصميم هذه البرمجية: وجـــود درجة معينة من التجاذب بين كل كلمة والكلمات المحيطة بها. ومن هنا فإن هذه االبرمجية تهدف إلى استخراج جميع التصاحبات اللغوية التي يكون بينها وبين الكلمة موضع البحث قوة جذب.

١- طريقة الحساب

المدخلات:

١) جدول تكراري للكلمات داخل ذخيرة واسعة النطاق.

٢) ملف توافق وقوع عن الكلمات المصاحبة للكلمة موضع البحث يتم استخراجه من الذخيرة.

المخرجات:

ملف توافق وقوع يتم ترتيبه على أساس درجة القيم المطلقة لكل سطر من أسطر الملف المفهرس.

الخطوات:

إذا تجاوزت أي كلمة W_i في سياق أحد أسطر ملف الـــسياقات قيمـــة معينة، يتم الحساب وفقًا للمعادلة التالية:

$$x_i = \frac{P_s}{P_c} = \frac{freq_{span}(w_i)}{N_{span}} / \frac{freq_{corpus}(w_i)}{N_{corpus}}$$

. حيث تشير P_s إلى معدل التكرار النسبي لعدد مرات ظهور الكلمة w_i في سياق له طول محدد، أما P_c فتشير إلى معدل التكرار النسبي للكلمة w_i داخل الذخيرة بأكملها، أي قيمة تكرار الكلمة في الجدول التكراري للكلمات.

وبعد ذلك يتم استخدام قيمة زد $Z - score$ في عمل تصنيف للنتيجة كما يلي:

$$\bar{x} = \sum_{i=1}^{n} \frac{1}{n} x_i$$

حيث تشير "n" إلى قيمة محددة تزيد عن حساب نتيجة "z" لجميع الكلمات التي تتواتر معًا. ولحساب قيمة z لكل كلمتين متصاحبتين يتم تطبيق المعادلة التالية:

$$z = \frac{x_1 - \bar{x}}{s}$$

حيث تشير "s" إلى قيمة الانحراف المعياري ويتم حسابها من المعادلة التالية:

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

وفي النهاية يتم جمع قيمة "z" التي تنتج من حساب درجة التصاحب بين جميع الكلمات الواردة في السجل المُفَهْرَس مع الكلمة موضع البحث، فنحصل على قيمة رقمية عن هذا السجل يُطلق عليها القيمة القياسية لهذا

السجل. ثم تُرتَّب السجلات تنازليًا حسب القيمة القياسية لتصاحب الكلمات الواردة في السجل مع الكلمة موضع البحث. ويكون شكل المخرجات التي تقدمها البرمجية عبارة عن قائمة مكوَّنة من مجموعة من السجلات يحتوي كلٌ منها الكلمات التي تتصاحب مع الكلمة موضع البحث، وأمام كل سجل تظهر القيمة القياسية لهذا السجل والسجلات المفهرسة التي تتمتع بقيم قياسية متساوية يتم ترتيبها أبجديًا.

وهنا نود الإشارة إلى أن القيمة القياسية لسجل مفهرس عبارة عن تجميع نتيجة z لكل تصاحب داخل هذا السجل، الأمر الذي يعكس ثقل كل عنصر من العناصر المكونة لهذا السجل. وإذا احتوى السجل المفهرس كلمة لها نتيجة "z" كبيرة، وكانت القيمة القياسية لهذا السجل مرتفعة، فإن السجلات المفهرسة الأخرى التي تحتوي كلمات لها قيمة "z" نفسها تكون لها القيمة القياسية نفسها. إلا أنه من عيوب استخدام الانحراف المعياري ما يلي: عندما يتساوى معدل تكرار بعض الكلمات مع متوسط عدد مرات ظهور هذه الكلمات فإن نتيجة "z" لهذه الكلمات تساوي صفرًا. ومن أجل التغلب على تلك المشكلة، قام سنكلير بحذف السجلات المفهرسة التي لها قيمة قياسية مرتفعة؛ التي تَرِد في مقدمة قائمة السجلات المفهرسة، وبعد ذلك قام بإعادة خطوات الحساب السابقة في استخراج قائمة سجلات مفهرسة جديدة لها قيمة قياسية مرتفعة.

٢- مثال

استَخدَم سنكلير وفريقه البحثي هذه البرمجية في عمل دراسة عن تصاحبات بعض الكلمات، ففي البداية تم تحديد كلمة "hot"، ككلمة لها عدد

من الدلالات، فمن الممكن أن تعبر عن ارتفاع درجة حرارة الجو، أو عن مذاق حَرِّيف، أو أكل طازج، وما إلى ذلك.

وفي هذا السياق تقوم برمجية تيبيكال **Typical** بترتيب ملف السياقات المستخرج لكلمة "**hot**" وفقًا للقيمة القياسية لكل سجل من سجلات الملف؛ حيث تقوم بوضع السجلات التي تحتوي تصاحبات متشابهة معًا على التوالي وفقًا لنتيجة "z". وتُجرِي البرمجية عملية البحث واستخراج سجلات ملف السياقات في سياق طوله ثلاث كلمات قبل وبعد الكلمة موضع الدراسة، على ألا تقل عدد مرات تكرار التصاحب داخل ملف السياقات عن سبع مرات. وفي النهاية تم استخراج ملف توافق وقوع كلمة "**hot**" بالإضافة إلى جدول معدلات التكرار لتصاحبات هذه الكلمة من ذخيرة إنجليزية تضم مائتي مليون كلمة. ويبين الشكل (٣-٣) التالي نتيجة البحث، علمًا أن الرقم الموجود في أول كل سجل يُشير إلى الرقم القياسي لهذا السجل، أما السجل المُفَهْرَس نفسه فيتم وضعه بين العلامتين <>.

19476.18 <asackful of guitas shaped like red hot pokers that stab the songs through>

19476.18 < a sackful of guitars shaped like red hot pokers! Visual fireworks: STEVE>

19476.18 < paims and pampas grass among the red hot pokers seem like a feasible and>

19474.55<like lupins and delphiniums, red hot pokers (Kniphofias), mullein>

19474.55< had gardens with raspberries and red hot pokers. Once we spent a week in the>

19474.55 <their weed-choked snapdragons and red hot pokers. If they ask about it,smile>

15446.13 <doctors found it helped dry vagina, hot flushes,sweats, ftension, anxiety and>

15082.65 <Problems of the menopause such as hot flushes,night sweats, dry vagina>

15076.28 <<FCH>sympyoms, which include hot flushes,sweats, tingling, and>

15070.74 <in particular can help with hot fushes, night sweats, vaginal>

... ...

13807.29 <an urgent need for the bathroom, hot and cold flushes and pins and>

13790.39 <turn white and that know what I mean. Hot and cold flushes and that> <M01> Was>

13790.39 <and pins and needles <FCH><M38><FCH> hot and cold flushes, sweating,>

13579.00 <service was held on a blisteringly hot day. When the body was lifted from>

13574.27 <<LTH>Sunday was blisteringly hot with cars and drivers alone>

13539.29 <designed to combat such blisteringly hot days whereas Malcolm roars like an>

13517.17 <said: We have enjoyed a blisteringly hot June, along with Scandinavia, the>

... ...

9875.08 <in. This is especially a danger in hot and hummid climates, such as the>

9866.32 <known as'dropouts', when played in hot and humid climates><LTH>Examination>

9676.71 <Phil found himself slaving over a hot grill at minimum wage while being>

9027.15 <Summers can be unbearably hot and humid and the scenery is flat>

8424.17 <<CQ1> I don't like slaving over a hot stove cooking a good meal and>

8341.5 <we spent hours lovingly slaving over a hot stove making, it's hardly surprising>

8334.75 <but when you are slaving over a hot stove in the kitchens of the Hotel>

8334.75 <<t>WOMEN spend hours slaving over a hot stove in the kitchen but are banned>

... ...

8135.44 <driving the criminals ' favourite hot hatchbacks cars # who face increases>

7822.09 <rising insurance costs,even more 'hot hatchbacks' like the new Citroen ZX>

7820.46 <aming the new Classics are the hot hatchbacks and homologation specials>

7820.46 <will go straight into group 20, and hot hatchbacks can expect to see their>

7820.46 <born by those owning what they call hot hachbacks and sports cars. Er the>

7816.97<on 45 high-risk models, mostly hot hatchbacks, was swift. New Escort RS>

7356.70 <display, and where summers can be hot and humid, sow in September where>

7120.72 <sheets (The dough is soft, so in very hot or humid weather, refrigrerate it for>

7174.93 <tin.<t> Baking Note: <FCH> In very hot or humid weather, or if your kitchen>

7037.53 <a day four days in weather so hot and humid that several men died. He>

7035.44 <and miantain a humid atmosphere in hot weather. Keep it shaded form the>

7020.37 <is partly affected by the weather. On hot humid days large amounts of pollen>

7005.79 <attack even without exercise. Very hot or humid weather will make it>

5465.49 <down because you stay dry and have a hot toddy when you get home>

5462.31 <a sherry at the theatre bar, or a 'hot toddy' to keep the cold at bay.>

5448.56 <He dipped his tiny beak into a hot toddy after this series of knight's>

5448.26 <you head off down the slopes- it's a hot toddy with an extremely potent kick>

5440.70 <but the Club were as cheering as a hot toddy> Even though they seemed>

5414.71 <leader Paddy Ashdown, branded the 'hot toddy' budget as a cynical manoeuvre>

شكل (٣-٣): نتيجة الاستعلام عن كلمة "hot"

مُفَهْرَسة تنازليًا حسب الرقم القياسي

يقدم الشكل (٣-٣) نتيجة الفهرسة لكلمة "hot" وفقًا للقيمة القياسية لكل سجل. ومن هذا الشكل يمكننا ملاحظة أن أكثر تصاحبات تحتوي كلمة "hot" هي "hot flushes"، و"hot and humid" وما إلى ذلك. ويمكننا أن نلاحظ كذلك أن أي تغير كبير في القيمة القياسية لسجلات الفهرسة يشير إلى استخدام جديد لكلمة "hot". ولذلك فإن التغير الذي يحدث في القيمة القياسية لسطر الفهرسة يمكن أن يشير إلى بداية ظهور استخدام جديد للكلمة موضع البحث.

٣- أثر مجموعة المتغيرات التي تحتاجها البرمجيتان على عمليات التحليل اللغوي

قبل استخدام البرمجيتين الإحصائيتين سالفتي الذكر، يُطلَب من المستخدم إدخال عدد من المتغيرات اللازمة للبرمجية، مع تحديد اسم البرمجية المستخدمة (Collocate أم Typical)، وبعد ذلك يقوم المستخدم بتحديد الكلمة موضع الدراسة وعدد السجلات التي سيحتويها ملف الفهرسة؛ التي سيتم حساب النتائج على أساسها؛ ذلك لأن النتائج تختلف تبعًا لاختلاف حجم الملفات التي يَجرى التحليل عليها. هذا بالإضافة إلى ضرورة تحديد طول النص قبل الكلمة موضع البحث وبعدها، وأقل قيمة لعدد مرات ظهور التصاحبات في إطار هذا الطول. وفيما يلي نعرض بعض الخبرات التي اكتسبها سنكلير من استخدام هاتين البرمجيتين.

(١) حجم ملف المدخلات

ليس هناك حدود تتطلبها برمجيتا الإحصاء تجاه حجم نص الفهرسة، فكلما زاد حجم نطاق ملف الفهرسة زادت مصداقية النتائج. وتشير الجداول

(٣-٤)، و(٣-٥)، و(٣-٦) إلى النتائج الإحصائية للتصاحبات؛ التي تم التوصل إليها عن طريق استخدام برمجية كولوكيت "Collocate" باستخدام ملفات فهرسة تحتوي ٥٠٠٠، ٢٠٠٠٠، ٥٠٠٠٠ سجل على التوالي، ففي ذلك الوقت تم اختيار سياق يصل إلى أربع كلمات، على أن يتم اختيار التصاحبات التي لا يقل عدد مرات تكرارها عن ١٠ مرات.

جدول (٣-٤): نتيجة التحليل على عدد ٥٠٠٠ سجل مفهرس

الكلمات المتصاحبة مع كلمة "hot"	عدد مرات ظهور الكلمة المتصاحبة في الذخيرة	قيمة وضوح درجة التصاحب	عدد مرات ظهور الكلمة المتصاحبة داخل ملف السياقات
Contorted	161	0.039	11
Flushed	691	0.167	11
Grin	1329	0.321	19
Mask	2221	0.536	26
Fines	1260	0.304	11
Starvation	1421	0.343	12
Smile	7371	1.780	54
Staring	2677	0.646	19
Brave	3549	0.857	22
Pale	4456	1.076	27
Expression	5760	1.391	31
Smiling	3123	0.754	14
Handsome	3263	0.788	13
Charges	15155	3.660	60
Face	49108	11.860	185
Tears	5429	1.311	19
Buried	4079	0.985	13
Neck	7236	1.747	22
Touched	3993	0.964	11
Prospect	6612	1.597	17
Value	21732	5.248	56
Thin	7341	1.773	17

جدول (٣-٥): نتيجة التحليل على عدد ٢٠٠٠٠ سجل مفهرس

الكلمات المتصاحبة مع كلمة "hot"	عدد مرات ظهور الكلمة المتصاحبة في الذخيرة	قيمة وضوح درجة التصاحب	عدد مرات ظهور الكلمة المتصاحبة داخل ملف السياقات
Volte	69	0.067	23
Expressio nless	165	0.159	29
Impassive	171	0.165	24
Contorted	161	0.156	22
Sallow	117	0.113	12
Ashen	120	0.116	12
Adversity	392	0.379	34
Creased	163	0.157	14
Flushed	691	0.668	51
Craggy	167	0.161	11
Frown	415	0.401	24
Haggard	260	0.251	15
Shadowed	234	0.226	12
Slap	910	0.879	46
Tanned	499	0.482	24
Slapped	764	0.738	36
Bony	308	0.298	14
Brightened	282	0.272	12
Streaked	284	0.274	12
Ruddy	285	0.275	11
Screwed	634	0.612	23
Beaming	420	0.406	15

جدول ٣-٦ نتيجة التحليل على عدد ٥٠٠٠٠ سجل مفهرس

عدد مرات ظهور الكلمة المتصاحبة داخل ملف السياقات	قيمة وضوح درجة التصاحب	عدد مرات ظهور الكلمة المتصاحبة في الذخيرة	الكلمات المتصاحبة مع كلمة "hot"
53	0.167	69	Volte
18	0.082	34	Barroom
73	0.398	165	Expressionles s
13	0.072	30	Eiger
60	0.389	161	Contorted
56	0.413	171	Impassive
24	0.215	89	Broderick
28	0.290	120	Ashen
20	0.215	89	Puckered
25	0.268	111	Reddened
13	0.142	59	Blotchy
22	0.251	104	Freckled
32	0.394	163	Creased
12	0.155	64	Redder
29	0.389	161	Puffy
66	0.947	392	Adversity
112	1.669	691	Flushed
19	0.283	117	Sallow
149	2.198	910	Slap
11	0.167	69	Sunburned
23	0.411	170	Smirk
22	0.403	167	Craggy

ونلاحظ من الجداول الثلاثة السابقة أن التصاحبات التي تم استخراجها من ملف الفهرسة الأكبر حجما لهـــا مصداقية أعلى.

(٢) طول السياق

يُعْتَبَر طول السياق هو أحد المتغيرات التـــي يُطلَـــب مـــن المـــستخدم تحديدها. غالبًا ما يتم تحديد السياق في اللغة الإنجليزية بأربع كلمـــات قبـــل

الكلمة موضع البحث وبعدها. فالسياق المصاحب للكلمة هو أحد مؤشـــرات المعنى لهذه الكلمة. فإذا كان لدينا كلمة متعددة الدلالة، فإن السياقات التي ترد فيها هذه الكلمة من الممكن أن تُعَبِّر عن معاني تلك الكلمة في حالاتها الدلالية المختلفة. ولذلك فإن هناك ضرورة لتحديد السياق المؤثر في الكلمة والـــذي يشير إلى عدد الكلمات التي ترد قبل وبعد الكلمة موضع الدراسة؛ وذلك من أجل اختيار طول السياق الذي يُقَدِّم أفضل نتيجة للتحليل.

فإذا استخدمنا برمجية كولوكيت **Collocate** في دراسة كلمـــة معينـــة، وفي كل مرة من مرات الاستخدام نقوم بتغيير طول السياق، ونحتفظ ببـــاقي المتغيرات دون تغيير، سنجد أن المخرجات التي تقدمها البرمجية تختلف في كل مرة. ويشير الجدولان (٣-٧)، و(٣-٨) إلى نتيجة العمل علـــى كلمـــة "eye"؛ وذلك من خلال ملف فهرسة طوله ٥٠٠٠ سجل، على ألا تقل عـــدد مرات تكرار التصاحب عن ١٠ مرات؛ ولكن مع تغيير السياق؛ حيث يشير الجدول الأول إلى النتيجة عندما يكون طول السياق كلمتين قبل كلمة "eye" وبعدها، والجدول الثاني عندما يكون السياق ٦ كلمـــات قبـــل كلمـــة "eye" وبعدها. وفيما يلي نعرض نتيجة تشغيل البرمجية على السياقين المختلفين.

جدول (٣-٧): نتيجة التشغيل على سياق طولـــه كلمتين قبل كلمة "eye" وبعدها

الكلمات المتصاحبة مع كلمة "eye"	عدد مرات ظهور الكلمة المتصاحبة في الذخيرة	قيمة وضوح درجة التصاحب	عدد مرات ظهور الكلمة المتصاحبة داخل ملف السياقات
Beady	99	0.012	12
Watchful	346	0.042	35
Remover	174	0.021	14
Untrained	249	0.030	13
Sockets	296	0.036	12
Socket	464	0.056	12
Blind	4941	0.097	81
Catches	1607	0.194	21
Naked	3486	0.421	40
Gel	1054	0.127	12
Caught	14201	1.715	133
Eagle	1912	0.231	18
Witnesses	3304	0.399	25
Eye	16359	1.975	121
Keeping	12294	1.485	76
Meets	5019	0.606	29
Contact	16184	1.954	90
Patch	2173	0.262	11
Witness	4342	0.524	21
Catching	2510	0.303	12
Keep	48681	5.878	211
Catch	8595	1.038	36

جدول ٣-٨ نتيجة التشغيل على سياق طوله ست

كلمات قبل كلمة "eye" وبعدها

عدد مرات ظهور الكلمة المتصاحبة داخل ملف السياقات	قيمة وضوح درجة التصاحب	عدد مرات ظهور الكلمة المتصاحبة في الذخيرة	الكلمات المتصاحبة مع كلمة "eye"
24	0.043	119	Beholder
13	0.036	99	Beady
20	0.063	174	Remover
36	0.125	346	Watchful
21	0.087	241	Glint
18	0.090	249	Twinkle
13	0.071	195	Contour
14	0.080	221	Retina
13	0.090	249	Untrained
14	0.107	295	Glam
12	0.107	296	Sockets
15	0.144	397	Blink
16	0.168	464	Socket
20	0.382	1054	Gel
90	1.790	4941	Blind
26	0.582	1607	Catches
242	5.926	16359	Eye
12	0.299	826	Makeup
43	1.263	3486	Naked
154	5.144	14201	Caught
18	4.454	12294	Keeping
18	0.693	1912	Eagle

ويمكننا أن نلاحظ أن الكلمات المتصاحبة مع كلمة "eye" مثل "beady"، و"watchful"، و"remover" وغيرها لها نسبة ظهور واضحة في

الجدولين، فضلاً عن أن بعض الكلمات التي لها درجة تصاحب عالية في الجدول (٣-٧) (مثل الكلمات "patch"، و"witness") لا تظهر في الجدول (٣-٨)، أو أن لها درجة تصاحب ضعيفة. والسبب في ذلك يرجع إلى كون الكلمات التي لها درجة تصاحب عالية من خلال السياق الذي يبلغ طوله كلمتين ستظهر في السياق الذي يبلغ طوله ست كلمات. ومن هنا، فإن درجة التصاحب الواضحة لهذه الكلمات في السياق الأقل سوف تضْعُف؛ بسبب تأثير الكلمات الموجودة في السياق الأكبر في نتيجة المعادلة. هذا بالإضافة إلى أن الكلمات التي لها درجة تصاحب أعلى، أو تلك التي تتمتع بحرية سياقية كبيرة سيتم الاحتفاظ بها في جدول التصاحبات المُستَخرَج من السياق الذي يبلغ طوله ست كلمات؛ وذلك مثل كلمة "patch" التي لا تظهر إلا بعد كلمة "eye"؛ ولذلك فقد ضعفت قيمة تصاحبها مع كلمة "eye" بسبب الكلمات التي تظهر دائمًا مع كلمة "eye" في سياق الست كلمات. وفي الجدول (٣-٨) نجد أن كلمة "beholder" لها درجة تصاحب عالية مع كلمة "eye"؛ ولكنها لم تظهر في الجدول ٣-٧؛ وذلك بسبب أن كلمة "holderbe" تُستَخدم دائمًا في التعبيرة "in the eye of the beeholder". ففي هذه الحالة نلاحظ أن كلمة "beholder" ظهرت خارج إطار السياق المُحَدَّد بكلمتين مع كلمة "eye". وحتى يمكن التغلب على هذه الظاهرة، يمكننا إضافة معلومة تحدد موضع كلمة التصاحب من الكلمة موضع البحث عند استخدام البرمجية، على سبيل المثال الاكتفاء بالكلمات التي تأتي بعد الكلمة موضع البحث أو قبلها، ففي تلك الحالة سترتفع درجة وضوح التصاحب.

(٣) الحد الأدنى لعدد مرات ظهور التصاحب

إن الحد الأدنى لعدد مرات ظهور التصاحب هو تحديد أقل قيمة لعدد مرات ظهور التصاحبات التي تتم عليها عملية الحساب داخل ملف السياقات. وقيمة هذا المتغير تتحكم بشكل مباشر في عدد التصاحبات التي تقوم البرمجية بانتخابها. فإذا كانت حدود هذه القيمة ضعيفة، فسوف تستغرق البرمجية وقت تشغيل أطول، وقد ينتج عن عملية التشغيل وجود بعض الأخطاء في النتائج؛ وعلى العكس من ذلك، إذا كانت تلك القيمة كبيرة، فسوف يؤدي ذلك إلى إهمال التصاحبات التي لها نسبة ظهور واضحة.

إن الغرض من تحديد عدد مرات ظهور التصاحب بأقل قيمة هو استبعاد الكلمات التي بها خطأ في الكتابة أو أسماء الأعلام وغير ذلك من الكلمات التي تظهر مرة أو مرتين داخل الذخيرة. ويمكننا أن نلاحظ من الجدول رقم ٣-٩ أهمية تحديد عدد مرات ظهور التصاحب بأقل قيمة. ويمكننا أن نكتشف من الجدول أن الكلمات الأكثر تمثيلاً للقوة التصاحبية مع كلمة "hard" قد ظهرت مرتين في مجمل الذخيرة، وعلى الرغم من أنها ظهرت مرة واحدة في تصاحب مع كلمة "hard"، فإن هذه الكلمات تتمتع بدرجة تصاحب عالية مع الكلمة موضع البحث. ونلاحظ من الجدول أيضًا أن كلمة "getthem" بها خطأ في الكتابة؛ حيث ينقصها مسافة فارغة في الوسط، والكتابة الصحيحة لها هي "get them".

جدول (٣-٩): تـــأثير تحديـــد عـــدد مـــرات ظهـــور التصاحب بأقل قيمة في نتيجة البرمجية

الكلمات المتصاحبة مع كلمة "hard"	عدد مرات ظهور الكلمة المتصاحِبة في الذخيرة	قيمة وضوح درجة التصاحُب	عدد مرات ظهور الكلمة المتصاحِبة داخل ملف السياقات
Anie	2	٠,٠٠٠	1
Bogna	2	٠,٠٠٠	1
Endochorion	2	٠,٠٠٠	1
Getthem	2	٠,٠٠٠	1
Givada	2	٠,٠٠٠	1
Hipp	2	٠,٠٠٠	1
Kinjiro	2	٠,٠٠٠	1
Korbel	2	٠,٠٠٠	1
Leinoff	2	٠,٠٠٠	1
Lektropaks	2	٠,٠٠٠	1
Leshenka	2	٠,٠٠٠	1
Maternite	2	٠,٠٠٠	1
Mogulled	2	٠,٠٠٠	1
Pittesburg	2	٠,٠٠٠	1
Pogrebnjak	2	٠,٠٠٠	1
Sarit	2	٠,٠٠٠	1
Shirtlifter	2	٠,٠٠٠	1
Spener	2	٠,٠٠٠	1
Tolars	2	٠,٠٠٠	1
Trancepotter	2	٠,٠٠٠	1
Weasling	2	٠,٠٠٠	1

وبصورة عامة، فإن قيمة المتغير الذي يتم إدخاله، ينبغي أن تتغير تبعًا لتغير طول السياق. فعندما يكون طول السياق قصيرًا، تكون تلك القيمة قليلة؛ وتزداد هذه القيمة تدريجيًا مع زيادة طول السياق.

الفصل الرابع

ترميز الذخائر اللغوية

إن تجميع عدد كبير من النصوص في لغة ما وتخزينها في الحاسب الآلي. يُكَوِّن ذخيرة لغوية هائلة الحجم. وبعد عملية التجميع، يسعى الباحثون إلى استخراج المعلومات التي يحتاجون إليها من هذه الذخيرة؛ على سبيل المثال إنشاء مُعْجَم لغوي أكثر جودة، أو دليل نحوي يعتمد على بيانات لغوية واقعية؛ وذلك من أجل فهم اللغة واستخدامها بشكل سليم وناجح. ولكي يمكننا استخراج المعلومات من الذخائر اللغوية المختلفة، فمن الضروري البدء بعمل تحليل للذخيرة اللغوية على مستوى واحد أو على عدة مستويات، بالإضافة إلى ترميز العناصر التي تنتج عن عملية التحليل داخل متن الذخيرة اللغوية، ومن ثَمَّ إعطاء قيمة مضافة أعلى إلى الذخيرة. وهذا ما يُطْلَق عليه ترميز الذخائر اللغوية. إن عمليات ترميز الذخائر اللغوية تم قبولها على نطاق واسع باعتبارها المدخل الأساسي للتعامل مع الذخائر اللغوية. ويناقش المرجع رقم [٥٨] بالتفصيل منهجيات ترميز الذخائر اللغوية على عدة مستويات لغوية، ونظرًا إلى أن سلسلة الكتب الأخرى التي نقدمها في هذا المجال بها شرح لمنهجيات ترميز الذخائر اللغوية الصينية، فلن نتناول في هذا الجزء طريقة الحساب المتعلقة بترميز الذخائر اللغوية، وسنكتفي فقط بتناول الموضوع من حيث معنى الترميز، ونماذج الترميز وأنواعه.

أولاً: معنى ترميز الذخائر

يمكن تعريف ترميز الذخائر اللغوية كما يلي: هو نوع من التطبيق يهدف إلى إضافة معلومات لغوية ومعلومات توضيحية للذخائر الشفهية أو التحريرية. ومن الممكن أن تشير كلمة "ترميز" إلى المنتج النهائي لهذه العملية: بمعنى العلامات اللغوية الملحقة بالذخائر اللغوية أو التي تتناثر وسط الذخائر اللغوية. ويُقصد بتلك العلامات اللغوية علامات المقاطع الصوتية، أو العلامات النحوية، أو الدلالية وما إلى ذلك. ونظرًا إلى أن اللغة الصينية المكتوبة لا تحتوي مسافات فاصلة بين حدود الكلمات، فإن عملية ترميز الذخائر اللغوية الصينية تشمل مرحلة إضافية تهتم بترميز حدود الكلمات، وعادة ما يُستَخدَم لهذه المرحلة من الترميز مسافة فارغة بين حدود كل كلمتين. إن أشهر مثال في عمليات ترميز الذخائر اللغوية وأكثرها قياسية هو الترميز النحوي، ويُطلَق عليه أيضًا ترميز الأنواع النحوية للكلمات أو ترميز POS. ففي عملية الترميز هذه يتم إضافة علامة أمام كل كلمة من كلمات الذخيرة من أجل تحديد نوعها النحوي. فعلى سبيل المثال "**公布/vgn**"، حيث نجد أن علامة "vgn" تشير إلى أن الفعل "公布" ويعني "أعلَنَ" فعل متعدٍّ لمفعول اسمي.

والسبب في قولنا: إن الترميز يحمل معلومات توضيحية هو أن الترميز في أقل تقدير عبارة عن مُنتَج يحمل فهم الإنسان للنص موضع الترميز. فترميز النصوص الصينية من حيث حدود الكلمات يعني إمكانية

تمييز الكلمات من بين سلاسل الرموز المتتالية في النصوص؛ التي لا يفصل بينها فاصل، بالإضافة إلى وضع ترميز نحوي لتلك الكلمات يوضح نوعها النحوي داخل النص وما إلى ذلك، بالإضافة إلى ذلك فإن هناك فرقًا بين كل من "الترميز"، و"الإيضاح" بالنسبة إلى النصوص. ويمكننا أن نُمَيِّز بين هذين النوعين من المعلومات من خلال النصوص التحريرية. فالعلامات الصرفية لنص تحريري يستخدم لها الحروف اللاتينية بما في ذلك علامات الترقيم والمسافات. وهذه العلامات يتم التعبير عنها رقميًّا بعلامات خاصة داخل الحاسب الآلي؛ حيث تشير كل علامة إلى شكل واحد تعبر عنه، بمعنى أن العلامات الصرفية الأصلية في النص يقابلها ملف رقمي واحد يعبر عنها. وفي أثناء عملية التمثيل هذه قد يفقد النص الأصلي بعض المعلومات التحريرية، مثل نوع الخط وحجمه. وهذا أمر مسموح به نظرًا إلى أن هذا النوع من المعلومات لا يُعَبِّر عن جوهر المعلومات اللغوية التي يعبر عنها النص. وعلى العكس من ذلك، فإن ترميز أي نص عبارة عن وحدة ما وراء لغوية (metalinguistic)؛ أي أن ما تقدمه عملية الترميز ما هو إلا معلومات لغوية عن ذلك النص، وليس مضمون النص نفسه.

إلا أنه بالنسبة إلى حوار شفهي، أحيانًا ما يكون هناك صعوبة في التمييز بين المعلومات الخاصة بالمضمون وتلك التي تختص بالإيضاح. وفي أثناء تحويل اللغة الشفهية إلى لغة تحريرية أو إلى نص رقمي، يضطر الشخص القائم بالتحويل إلى الجمع بين نظامي العلامات المختص بالمضمون والآخر المختص بالإيضاح. وفي معظم عمليات التحويل، يتم استخدام

الوصف الصوتي بشكل إضافي لمزيد من المساعدة وذلك بالإضافة إلى الكتابة الأبجدية؛ ولكن هذا التدخل يتيح إمكانية القراءة الصوتية بشكل سطحي، أما فيما يتعلق بالطبيعة المادية أو اللغوية أو الاجتماعية للنص فهذا مستوى مُعَقَّد بدرجة كبيرة ومن الصعب تحقيقه. على سبيل المثال عند ترميز مستويات التأكيد والتنغيم الصوتي في اللغة الشفهية، غالبًا ما يتم ذلك عن طريق الحكم الشخصي للقائم بعملية التحويل من اللغة الشفهية إلى اللغة التحريرية، وفي الوقت نفسه يتوقف ذلك على نظام التحليل المُسْتَخْدَم.

ثانيًا: لماذا يتم ترميز الذخائر اللغوية؟

١- استخراج المعلومات

لا يمكن اعتبار الذخائر اللغوية ذات فائدة إلا إذا أمكن استخراج المعلومات والمعارف منها. وفي الحقيقة فإن استخراج المعلومات من الذخائر اللغوية يُحَتِّم من البداية غرس بعض المعلومات الإضافية، والمقصود بذلك إضافة علامات الترميز. فالذخائر اللغوية التي تمثل نصوصًا رقمية إذا لم تُجْرَ عليها أية عملية من عمليات المعالجة يُطْلَق عليها ذخيرة لغوية خام (**raw corpus**)، فمثل هذه الذخائر وعلى الأخص الذخائر الخام الصينية ينقصها المعلومات الصرفية والنحوية وما إلى ذلك، مما يجعل قيمتها ضئيلة إلى حدٍّ كبير. على سبيل المثال: كلمة "**left**" في اللغة الإنجليزية باعتبارها كلمة مضادة في المعنى لكلمة "**right**"، كما في "**my left hand**"؛ وفي الوقت نفسه يمكن لهذه الكلمة أن تأتي ظرف مكان فنقول"**turn left**"، أو اسمًا فنقول

"on your left". ولكن إذا تم النظر إليها على أنها صيغة الماضي من الفعل "leave" سيتم تحليلها باعتبارها فعلاً، مثل "I left early" ومن هنا فإن كلمة "left" لها عدد من الاستخدامات. إلا أن هذا التعدد في المعنى لهذه الكلمة لا يمكن تمييزه بوضوح في ذخيرة لغوية خام لم يتم عليها أية عملية من عمليات الترميز. وهذا النوع من الذخائر سيكون به قصور شديد إذا تم التعامل معه كمصدر من مصادر صناعة المعاجم. أما إذا تمت عملية الترميز النحوي للذخيرة، فإن كل مرة تظهر فيها كلمة "left" سيكون أمامها رمز يوضح نوعها النحوي، وهذه المعلومات من شأنها أن تساعد في تحسين العمل المعجمي. ومثال آخر في مجال التطبيقات الخاصة بتحويل النصوص التحريرية إلى نصوص منطوقة (Text To Speech)، نجد أن كلمة lead في اللغة الإنجليزية عندما تكون اسمًا تُنْطَق /led/، وعندما تكون فعلاً تُنْطَق /li : d/. فإذا كنا بصدد تطوير برمجية ناطقة (بمعنى تحويل مُدْخلات الحاسب الآلي من نصوص تحريرية رقمية إلى مُخرَجات صوتية) فإن هذه البرمجية سيتحتم عليها أن تُمَيِّز ما إذا كانت كلمة lead اسمًا أم فعلاً قبل أن تنطقها نطقًا سليمًا. فضلاً عن أن اللغة الصينية تنتشر بها ظاهرة الرموز التي لها أكثر من نطق، مثل: الرمز "行" الذي يُنطَق "hang2" كما في "银行" بمعنى بنك، وينطق "ingx2" كما في كلمة "行人" بمعنى المارّة. في ذلك الوقت وعند القراءة الصوتية للنصوص التحريرية ستكون هناك حاجة لترميز الرمز "行" حتى تتمكن الآلة من نطقه بشكل سليم. ومن هنا فإن عمل ترميز صرفي ونحوي للذخائر اللغوية من شأنه أن يُقَدِّم المعلومات التي تحتاجها تلك البرمجية الناطقة.

٢- تعدد الاستخدام

إن المقصود بتعدد استخدام الذخائر اللغوية التي تحمل علامات الترميز هو إمكانية إعادة استخدام المصادر اللغوية التي تقدمها تلك الذخائر غير مرة. ويعتقد البعض أنه لا حاجة إلى إضاعة الوقت والمجهود في أعمال ترميز شاملة للذخائر اللغوية، وأنه يمكن الاكتفاء بتصميم برمجية ذكية تقوم بتمييز الأنواع النحوية للكلمات؛ مثل كلمة "left" حيث يتم التعامل معها باعتبارها صفة إذا جاءت قبل اسم، ويتم التعامل معها على أنها ظرف مكان إذا جاءت بعد فعل، وهكذا. ولكن تنفيذ الأمر على هذا النحو يشوبه عيبان:

(١) يشير المثال السابق إلى أننا إذا أردنا تمييز إحدى الكلمات، فمن الضروري أن نتعرف سابقًا على نوع الكلمة التي تسبقها. ولذلك فإن تمييز نوع الكلمات لا يمكن النظر إليه بمعزل عن الكلمات المحيطة بالكلمة التي نرغب في ترميزها.

(٢) الهدف من عمل الترميز النحوي وغيره من مستويات الترميز للذخائر اللغوية هو: تحويل المادة اللغوية التي تحتويها الذخيرة إلى مادة ذات قيمة أعلى، وهذا الغرض يتحقق بمجرد إضافة مستويات الترميز المختلفة؛ حيث تتيح هذه العملية الذخيرة للآخرين كي يستفيدوا منها. إن عمليات ترميز الذخائر اللغوية مُكلِفة ومُستنفِذة للوقت؛ ولكن كل ذلك المال والجهد يكون في محله إذا أمكننا تنويع استخدامات الذخائر.

٣- تعدد الوظائف

إن الذخيرة التي تتم عليها عمليات الترميز غالبًا ما يكون لها أهـداف وتطبيقات مختلفة، وهذا ما يُطلَق عليه تعدد الوظائف للذخيرة اللغوية. ففـي الجزء السابق تحدثنا عن وظيفتين مختلفتين للذخائر اللغوية المُرَمَّزة وهمـا الاستخدام في صناعة المعاجم وإنتاج تطبيقات القراءة الجهريـة للنـصوص التحريرية. ويضاف إلى ذلك أن الذخائر المُرَمَّـزة يمكـن أن يَنْـتُج عنهـا تطبيقات أخرى في مجالات هندسة اللغة، مثل الترجمة بمـساعدة الحاسـب واستخراج المعلومات من النصوص، وما إلى ذلك. وهكذا، فـإن عمليـات الترميز من شأنها أن تحقق "قيمة مضافة" للذخائر اللغوية بكـل مـا تعنيـه الكلمة. والترميز النحوي، باعتباره أحد مستويات الترميز الأساسية، ما هـو إلا تمهيد للمستويات الأعلى من الترميز، على أساس أن هـذا النـوع مـن الترميز هو الخطوة الأولى نحو الترميز على مستوى بنية الجملة والترميـز على مستوى الدلالة. ونظرًا إلى وجود العديـد مـن المـستخدمين الـذين سيستفيدون من الذخائر اللغوية المُرَمَّزة، فسيأتي من ضمنهم من يتمكن مـن تفعيل استخدامات الترميز بشكل لم يَرِد في تصور مصممي عمليات الترميز من البداية؛ الأمر الذي يزيد من أهمية الأدوار الإضافية التي تلعبها الذخائر اللغوية المُرَمَّزة في مجال العمل اللغوي.

ثالثًا: التوحيد القياسي لعمليات ترميز الذخائر اللغوية

إن مستوى "الخبرة" الذي يتمتع بهَ القائمون على ترميز الذخائر اللغوية وقياسية علامات الترميز المستخدمة ومدى منطقيتها وقابليتها للاستخدام كلها

عناصر يتحدد على أساسها مدى فائدة المعلومات التي يُقَدِّمُها ترميز الذخيرة اللغوية وإمكانية احتوائها على معلومات يمكن التعامل معها. ففي الفترة القصيرة من تاريخ علم ترميز الذخائر اللغوية، كانت أنظمة الترميز التي يقدمها العاملون على بناء الذخائر اللغوية صعبة الاستخدام من قِبَل الآخرين لدرجة أن هناك بعض الحالات كان يستحيل استخدامها. وحتّى يمكن تجنب هذه الظاهرة، ينبغي تحقيق المعايير التالية في تصميم أنظمة الترميز:

١- إمكانية استعادة النسخة الخام للذخيرة اللغوية بعد حذف علامات الترميز، وبعبارة أخرى إمكانية استرجاع أصل الذخيرة.

٢- إمكانية استدعاء المعلومات التي يتم ترميزها بشكل مستقل عن الذخيرة، مع إمكانية حفظ النتيجة عند الحاجة إلى ذلك.

٣- إمكانية قيام مستخدم الذخيرة بالاطلاع على مستندات تحتوي المعلومات التالية:

أ) المنهجية المتبعة في الترميز، أي الحصول على مستندات تحتوي وصفًا وشرحًا للمعايير التي تم الاتفاق عليها في عمليات الترميز.

ب) مستندات تضم أسماء القائمين على عمليات الترميز والمكان الذي تمت فيه والمنهجية المُتَّبَعة.

ج) شرح تفصيلي لما يتعلق بالخطوات التي اتُّبِعَت لضمان جودة الترميز؛ وذلك نظرًا إلى كثرة حدوث الأخطاء واللبس الذي يحدث في أثناء عمليات الترميز بسبب عدم توحيد المعايير المُستَخدَمة؛

ويُذكَر على سبيل المثال: تسجيل المستوى الذي وصلت إليه عمليات مراجعة الترميز، والنسبة المئوية للأخطاء التي تم اكتشافها في أثناء عمليات المراجعة، والمستوى الذي وصلت إليه عمليات توحيد معايير الترميز، وما إلى ذلك.

د) وضع إشارة للمستخدم مُفادُها أن عمليات الترميز التي تمت على الذخيرة ليست مطلقة وأنه لا يمكن تجنب وجود قدر من الأخطاء، وأن الذخيرة لا تخرج عن نطاق أنها مصدر لغوي تتحقق الفائدة منه عند حد معين فقط. يقتصر الغرض من إتاحة معلومات عن المنهجية المتبعة في الترميز على تقديم مرجعية يمكن الاستشهاد بها والاستفادة منها وقت استخراج التطبيقات المختلفة من الذخيرة اللغوية. وفي تلك الحالة، سيكتشف العديد من المستخدمين أن هناك فائدة من استخدام الذخائر اللغوية المُرَمَّزة، وأن ذلك أفضل من اتباع منهجيات قاموا بتصميمها بأنفسهم؛ لأن هذا العمل يحتاج إلى إضاعة العديد من السنوات حتى يمكن إنجازه.

هـ) بناء منهجية للترميز، وحتى يمكن تجنب سوء الفهم وضمان حسن استخدام المتعاملين مع الذخيرة، تم الاعتماد على بيانات للتحليل تتصف بالوسطية ووجود أساس نظري لها بالإضافة إلى إمكانية الاتفاق عليها من أكبر عدد من الناس. وعلى الرغم من حتمية تعرض عمليات الترميز لبعض الخلافات النظرية، فإن الغرض الأساسي من عمليات الترميز يتجه بشكل أكبر إلى الالتزام قدر الإمكان بمدى القبول والفهم على نطاق واسع من المستخدمين.

و) أية منهجية للترميز لا يمكن أن ينتج عنها ما يسمى "بالمعايير التي لا تتغير". فقد ثبت من خلال التطبيق العملي أن منهجيات الترميز تتجه دائمًا إلى التغيُّر. على سبيل المثال، نجد أن نطاق الذخائر التي تم ترميزها قد يتسبب في إعاقة عمليات الترميز في المستويات العليا الأكثر تفصيلاً؛ لأن تحقيق الهدف الأساسي من عمليات الترميز يتطلب وضع أولوية للتفكير في بعض المعلومات التي تخص عددًا من التقسيمات التي تنشأ على أساسها منهجيات الترميز وما إلى ذلك.

على الرغم من توافر المبادئ الستة سالفة الذكر، فما زال هناك البعض يطالب بإقرار نوع من المعايير القياسية في مجال ترميز الذخائر اللغوية؛ بالإضافة إلى أن التطبيقات المختلفة في السنوات الأخيرة في هذا المجال قد توصلت تدريجيًّا إلى إيجاد نوع من التوحيد العلمي لعمليات الترميز. ويتمثل أحد أسباب التوجُّه إلى التوحيد المعياري في الوصول إلى مستوى الشيوع والتعميم؛ لأنه ما إن يجد المتعاملون في هذا المجال فائدة من استخدام أحد نماذج الترميز، حتى يُصِرُّون على استخدام هذا النموذج لتطوير ذخائرهم التي سبق ترميزها. وهناك سبب آخر وهو ما سبق التأكيد عليه من مبدأ تعدد الاستخدامات. فإذا رغب عدد من الباحثين تبادل البيانات أو المصادر اللغوية (ذخائر مُرَمَّزَة على سبيل المثال) فسوف يكون هذا التبادل سهلا بين الجهات المختلفة إذا ما كانت تتبع نموذجًا موحدًا للترميز أو على الأقل بينها اتفاق على بعض المبادئ الاسترشادية. وفي حالة الحاجة إلى

تبادل البرمجيات التي تتعامل مع الذخائر اللغوية فسوف تبرز قضية التوحيد القياسي لعمليات الترميز كمطلب أساسي لإمكانية تبادل تلك البرمجيات.

رابعًا: الحدود التي يتم عندها ترميز الذخائر اللغوية

هناك اختلاف شديد بين الحدود التي تقف عندها عمليات الترميز بين كل من اللغتين الإنجليزية والصينية. ويرجع السبب في ذلك إلى عدم وجود مسافات تشير إلى حدود الكلمات في اللغة الصينية. وتُعتَبر عملية تقسيم سلاسل الرموز في النصوص الصينية إلى سلاسل من الكلمات هي المهمة الأولى في عملية ترميز اللغة الصينية للتعرف على حدود الكلمات آليًّا. إن تمييز حدود الكلمات في اللغة الصينية هو أحد المشروعات الأساسية المهمة في علم معالجة اللغات الطبيعية باللغة الصينية؛ وتُحقق تلك الخطوة هدفًا أساسيًا في مجال معالجة النصوص آليًا؛ حيث تُعتَبَر الكلمات وحدات المعالجة الآلية للنصوص، وهذا ما ينبغي أن يتم كخطوة أولية يجب التغلب عليها بالنسبة إلى المنظومة الكاملة للمعالجة الآلية للغة الصينية. إن عمليات المعالجة الآلية للغة الصينية تشمل عمليات الاستعلام عن المعلومات واستخراجها، والترجمة الآلية، والتحليل النحوي وغير ذلك من العمليات اللغوية التي لا غنى عنها عند التعامل مع الكلمات باعتبارها الوحدات الأساسية للتعامل اللغوي. ومن خلال جهود العلماء على مدى عشرين عامًا حققت اللغة الصينية إنجازات كبيرة في مجال التمييز الآلي لحدود الكلمات الصينية، وقد وصلت نسبة دقة التمييز إلى حوالي ٩٩% [٥٩]. ولكن ما زالت

هناك بعض المشكلات في مجال تمييز حدود الكلمات داخل النصوص الصينية لم يتم حلها بشكل جذري. وتشمل تلك المشكلات: أسماء الأعلام الشخصية والجغرافية وأسماء الآلات وما إلى ذلك من الكلمات التي لم تُسَجَّل في قوائم الكلمات التي يتم التمييز على أساسها، بالإضافة إلى إشكاليات اللبس في تمييز حدود الكلمات.

وفي الجزء التالي نستعرض المراحل التي تمر بها عمليات الترميز؛ التي تم التعارف عليها على مستوى مختلف اللغات: الترميز النحوي (أو ترميز الأنواع النحوية للكلمات) (**grammatic tagging**)، والترميز على مستوى بنية الجملة (**syntactic annotation**)، والترميز الدلالي (**semantic annotation**)، والترميز على مستوى الخطاب (**discourse annotation**).

١- الترميز النحوي

كان ذلك هو أول مشروع لترميز ذخيرة لغوية، وقد تم على ذخيرة براون عام ١٩٧١ بجامعة براون الأميركية. وكان ذلك بإشراف عالِمَي لغة أشرفا على عملية الترميز هما فرانسيس **Francis** وكوسيرا **kucera**، وتنفيذ اثنين من باحثي الماجستير قاما باتباع القواعد المتعلقة بالسياق في تنفيذ هذا النوع من الترميز. وتضم مجموعة الترميز الخاصة بهذه العملية عدد سبعة وسبعين رمزًا. وهذه الرموز النحوية لا تُمَيِّز فقط الكلمات من حيث النوع النحوي مثل الأسماء والأفعال والصفات وما إلى ذلك، بل تصل إلى تصنيفات أكثر تفصيلاً داخل كل نوع، مثل تمييز صيغة المفرد والجمع من الأسماء، وتمييز الأنواع المختلفة للصفات، وما إلى ذلك.

وقد وصلت دقة النتائج التي حققتها برمجية الترميز النحوي تلك إلى ٧٧%، وقد تم تصويب أخطاء الترميز بالمجهود البشري بعد ذلك، وفي النهاية تم التوصل إلى مُنتَج ذي فائدة عالية، ألا وهو ذخيرة براون المُرمَّزة نحويًا. ويَرْجِع مغزى هذا العمل البحثي إلى أنه أظهر لأول مرة الخصائص العامة لترميز الذخائر اللغوية. فمن ناحية، أظهر هذا العمل الفرق بين العمل الآلي والعمل اليدوي في ترميز الذخائر اللغوية؛ حيث أكد حتمية أسلوب الترميز الآلي، مع ضرورة أن يَتبَع هذه المرحلة عمل يدوي مكثف ومجهود شاق في عمليات المراجعة والتصويب. فالترميز اليدوي والترميز الآلي عملان يكمل كل منهما الآخر، ولا يمكن بأي حال من الأحوال الاكتفاء بالترميز اليدوي فقط في إنجاز تلك المهمة. ومن ناحية أخرى فإن الترميز الآلي لا يمكن الاعتماد عليه إلا بعد أن يصل إلى درجة عالية من الدقة في الترميز.

والمشروع الثاني للترميز النحوي تم تنفيذه عام ١٩٨٢ على ذخيرة لوب **LOB**. يتمثل وجه الاختلاف بينه وبين المشروع السابق في: تطبيق منهجية الاحتمالات الإحصائية على الذخيرة اللغوية موضع الترميز. وقد اعتمد هذا المشروع على ذخيرة براون المُرَمَّزَة نحويًا كمصدر للإحصاءات اللغوية في حساب احتمال تحول علامتي ترميز نحويتين داخل ذخيرة لوب وغير ذلك من المتغيرات. وقد وصلت دقة برمجية الترميز النحوية تلك إلى ٩٦,٧% وأُطلِق عليها سم **CLAW1** ومعنى ذلك أنه بالمقارنة بالبرمجية الأولى التي اعتمدت على القواعد كمصادر للترميز، فإن دقة الترميز بينهما

قد قفزت قفزةً واحدةً بمقدار ٢٠%. بعد ذلك، ظهرت تباعًا العديد من برمجيات الترميز النحوي التي اعتمد معظمها على نظرية الاحتمالات الإحصائية. والمشكلة الوحيدة في تلك المنهجية هو احتياجها إلى ذخيرة لغوية سبق ترميزها للتجريب عليها، بالإضافة إلى أن طول السياق يكون محدودًا في أثناء حساب الاحتمالات، بمعنى ضرورة الاكتفاء بكلمة أو كلمتين يمين الكلمة موضع التعامل أو يسارها. وقد بدأت دراسات الترميز النحوي للغة الصينية في نهاية الثمانينيات وبداية التسعينيات من القرن العشرين، وكانت أول جهتين تقومان بعمل دراسة عن الترميز النحوي للغة الصينية هما جامعة شين خوا وجامعة شان شي.

٢- الترميز على مستوى أبنية الجُمَل

المقصود بالترميز على مستوى أبنية الجُمَل إضافة معلومات تبين الأبنية النحوية لجمل الذخيرة اللغوية. إن أول من طرح فكرة دراسة ترميز الذخائر اللغوية على مستوى بنية الجمل هو العالم إليجارد (**Ellegard**)[٦٠]؛ حيث قام هو وتلميذ له عام ١٩٧٨ بعمل تحليل نحوي لجزء من ذخيرة براون (حوالي ١٢٨٠٠٠ كلمة). ومع حلول ثمانينيات القرن العشرين، بدأت جامعة ناجميجن (Nijmegen)[٦١] وجامعة لانكاستر (lancaster)[٦٢] في إنشاء برمجية يمكنها إجراء تحليل نحوي لذخيرة لغوية. ومع بداية التسعينيات، أثبتت بنوك التحليل الشجري (Tree Banks) أن الذخائر المُرَمَّزة على مستوى أبنية الجمل هي أحد المصادر المهمة في مجال

المعالجة الآلية للغات الطبيعية. على سبيل المثال، في أثناء تمييز الأصـــوات اللغوية ومجال الترجمة بمساعدة الحاسب تكون هناك حاجة إلى برمجية قوية للتحليل على مستوى أبنية الجمـــل. ويـــصل بنـــك لانكاســـتر- آي بـــي إم (Lancaster/IBM) للتحليل الشجري إلى ثلاثة ملايين كلمة، أما إنشاء بنـــك جامعة بنسلفانيا[٦٣] فقد أضاف قطاعًا عريضًا من مستخدمي هذا النوع مـــن المصادر اللغوية؛ حيث استخدمت تشيلبا (Chelba) من جامعة جون هوبسكن بنك التحليل الشجري هذا في تأسيس نموذج لغـــوي يعتمـــد علـــى الأبنيـــة النحوية[64]، وحلّت بذلك مشكلة التصاحب اللغوي طويل المدى في النصوص اللغوية، وحققت التجارب المبدئية لهذا النموذج نسبة دقة عالية فـــي تمييـــز الجمل. ويشير مصطلح بنك التحليل الشجري إلى أن شجرة التعبيرات اللغوية هي الوحدة الأساسية في عمل ترميز الذخائر اللغوية على مـــستوى الأبنيـــة النحوية. إن مشروع الترميز على مستوى أبنية الجملة أكثـــر تعقيـــدًا ولـــه احتياجات أكثر تقدمًا من مستوى الترميز النحوي للكلمـــات. ولـــذلك تـــأتي الدراسات المتعلقة بهذا المجال في مرحلة تالية لمرحلـــة الترميـــز النحـــوي للكلمات، وإلا فشلت تلك المشاريع وكانت نتائجها في التحليل غير دقيقة.

وفي أثناء عمل ترميز للذخيرة اللغوية على مستوى أبنية الجمل يمكن عمل تحليل نحوي كلي أو جزئي لجمل الذخيرة. ومن خلال تحليل الـــذخائر اللغوية على مستوى أبنية الجمل يمكن التعمق في الاســـتفادة بالاســـتخدامات التالية للذخائر اللغوية:

(أ) ابتكار برمجيات لتحليل الذخائر على مستوى أبنية الجمل وتطويرها

إن الاستخدام الرئيس لعمليات ترميز الذخائر اللغوية على مستوى أبنية الجمل هو التدريب على تصميم برمجيات للترميز الآلي لهذا المستوى وتدريبها على الترميز الصحيح، هذا بالإضافة إلى أن هذه البرمجيات من العناصر المحورية التي لا غنى عنها في التطبيقات المتعلقة بالمعالجة الآلية للغات الطبيعية. إن استخدام الذخائر اللغوية المرمزة على هذا المستوى يساهم في تصميم برمجيات لتحليل الذخائر اللغوية على مستوى أبنية الجمل تعتمد على نظرية الاحتمالات، الأمر الذي يُدَعِّم إمكانات تلك البرمجيات في الترميز على هذا المستوى. وقد صَمَّمَ كل من جيلينيك (**Jelinek**)، وكولينز (**Collins**) من جامعة بنسلفانيا الأميركية برمجية تحليل لأبنية الجمل باستخدام نظرية الاحتمالات، وللتعرف عليها بالتفصيل يمكنك الاطلاع على المرجعين [65]، و[66].

(ب) استخراج المعلومات المتعلقة بالمفردات

الذخائر اللغوية المرمزة على مستوى أبنية الجمل تحتوي العديد من المعلومات الصرفية والنحوية؛ ومن ثَمَّ يكون لها فائدة في إنشاء المعاجم الإلكترونية. والمعاجم الإلكترونية عبارة عن مصادر ذات أبنية تُقَدِّم المعلومات اللازمة عن التغيرات الصرفية التي تحدث للكلمات بالإضافة إلى المعلومات النحوية والدلالية؛ وذلك لتغذية أنظمة المعالجة الآلية للغات

الطبيعية. إن استخدام هذا النوع من الذخائر اللغوية يمكنه أن يقدم للمعاجم الإلكترونية معلومات عن تصاحبات المفردات وأُطُر الاستخدام، بالإضافة إلى بعض المعلومات عن استخداماتها في أنواع النصوص المختلفة.

٣- الترميز على مستوى الدلالة

يتم استخراج المفردات التفصيلية لعملية الترميز الدلالي من المستويات اللغوية المختلفة. ففي البداية، يتم عمل ترميز دلالي لكل كلمة من كلمات النص، ويعتمد جوهر هذه العملية على تمييز التعدد الدلالي للكلمات وفقًا للسياق والتوصل إلى المعنى الصحيح لكل كلمة. ولذلك إذا أردنا التحدُّث بصورة أدق، فإن هذا المستوى من الترميز ينبغي أن يُطلَّق عليه اسم ترميز معاني الكلمات أو عملية إزالة اللبس الدلالي بين الكلمات، ويُطلَّق عليها اختصارًا WSD (Word Sense Disambiguation) بالإضافة إلى ذلك، يمكن ترميز كل جملة من جمل النص من حيث معنى الجملة؛ فعلى سبيل المثال، يُمكِن استخدام شبكة العلاقات الدلالية التي تنشأ من الحالات النحوية للكلمة داخل التركيب من أجل التعبير عن المعنى المنطقي للجملة، أو الاكتفاء باستخدام العناصر الدلالية الثلاثة التي تتكون من عنصرين لغويين مع الحالة الدلالية التي تربط بينهما للتعبير عن معنى كل وحدة دلالية تتكون منها الجملة. والجدير بالذكر أن الباحثين في مركز الأبحاث التابع لشركة مايكروسوفت الأميركية قد استخدموا رسميًّا العلاقات الدلالية الثلاث باعتبارها خلية لغوية وقاموا بتكوين شبكة دلالية هائلة الحجم أُطلِقَ عليها اسم

مايندنِت (MindNet)[٣٧] اعتمادًا على معجمين للغة الإنجليزية ونتائج التحليل النحوي والدلالي لإحدى الموسوعات باللغة الإنجليزيــة. وتُــستَخدَم شــبكة مايندنِت (MindNet) حاليًا في إزالة اللبس الدلالي والنحوي، بالإضافة إلــى استرجاع المعلومات وما إلى ذلك من الأبحاث المتعلقة بحقل المعالجة الآلية للغات الطبيعية ومشروعات تطوير هذا العلم.

وفي مجال تحليل الخطاب، فإن أسلوب التعبير عن المفاهيم من الممكن أن ينعكس على النصوص أو الشكل الإدراكي الذي يتكون بين المتحــدثين. ففي الحديث الذي يدور بين الطبيب والمريض، ســوف يــستخدم الطبيــب بالتأكيد كلمات من قبيل "منطقة البطن" وغيرها من التعبيرات الاصــطلاحية للتعبير عن مستوى معين من التخصص في الحديث، وطبيب آخر قد يستخدم كلمة أكثر عامية مثل كلمة "البطن" حتى يمكنه التواصل مع المريض حسب مستواه المعرفي. بالإضافة إلى ذلك ففي مجال البحث عن المعلومــات، إذا أراد شخص مهتم بالموضة أن يتعرف، من خلال ما تنشره الصحف، علــى التغيُّر الذي حدث في مجال الأزياء، واتخذ من ارتداء السِّروال مثالاً لــذلك، ففي أثناء تعامله مع ذخيرة لغوية سيكون من الطبيعي ألا يكتفي بالبحث عن كلمة سِروال فقط، بل ينبغي له إضافة كلمات أخرى تمثل أنواع الــسراويل مثل "السِّروال القصير"، و"السِّروال اللــصيق (الاســتريتش)"، و"الــسِروال الجينز"، و"سِروال سباق الخيل" وما إلى ذلك. وهذا ما نطلق عليه في علــم الدلالة إشكالية الكلمات المتعددة التي لها دلالة واحدة. وبمعنى آخــر وجــود عدد من الكلمات تشير إلى مفهوم واحد في الوقت نفسه. بالإضافة إلى ذلك،

نجد أنه في مجال البحث عن المعلومات تكون هناك حاجة إلى حل إشكالية الكلمات متعددة الدلالة. فإذا أردنا الآن أن نتعرف على مقدار التغير في مجال المواد، يمكننا أن نبحث بالكلمة المفتاحية "材料" "مواد"، إلا أن كلمة "材料" "مواد" تلك قد تحتمل معنى آخر هو "文档" "وثائق". وأهم ما في الأمر هو أن المستخدمين لا تكون لديهم رغبة إلا في الاستعلام عن المعلومات التي تهمهم فحسب.

وتُعتَبَر الحاجة لحل مثل هذه المشكلات إحدى آليات الترميز الدلالي للنصوص؛ بمعنى أن وضع علامة تشير إلى معنى كل كلمة داخل النص يُعتَبر بمثابة تقرير لمعنى هذه الكلمة وفقًا للسياق التي وردت فيه. والمثال السابق يوضح لنا أن الرمز المُستَخدَم في الترميز الدلالي يشير إلى نطاق دلالي معين تنتمي إليه الكلمة موضع الترميز وهذا النطاق الدلالي يضم في إطاره مجموعة أخرى من الكلمات.

وعند إجراء الترميز الدلالي، ينبغي البدء باختيار نظام التصنيف الدلالي (أو المفهومي) الذي سيتم اتباعه في عملية الترميز. وفي هذا الإطار، ينبغي النظر إلى العناصر التالية بعين الاعتبار:

(١) أن يكون نظام التصنيف الدلالي المُتَّبَع متعارَف عليه بين جمهور علماء اللغة أو علماء اللغة النفسيين.

(٢) أن يُغطِي هذا النظام الكلمات الحقيقية في إحدى اللغات وليس جزءًا منها.

(٣) أن يُمكِن تعديل هذا النظام بمرونة، حتى يسهل تطويعه لخدمة مستخدمين آخرين ومجالات أخرى.

(٤) أن يكون حجم الوحدات الدلالية المستخدمة مناسبًا.

(٥) أن يتمتع نظام التصنيف الدلالي بالتدرج البنائي.

(٦) أن يعتمد نظام التصنيف الدلالي على معيار قياسي موحَّد.

٤- ترميز مستوى الخطاب: الترميز على مستوى العلاقة بين المتحدثين

يختلف الترميز على مستوى الخطاب عن المستويات الأخرى من مستويات الترميز في صعوبة وضع تعريف محدد له. فعند ترميز معلومات الخطاب لنص من النصوص، يمكن الاستعانة بالجمل كوحدات للترميز، حيث يُنظَر إلى الجمل على أنها أكبر وحدة للتركيب النحوي، وعند ذلك الحد يتم ترميز الجملة، بالإضافة إلى تصنيف الجملة وفقًا لوظيفتها في عملية الخطاب؛ هذا بالإضافة إلى إمكانية استخدام مفاهيم مثل "محور الجملة"، و"موقع الفاعل"، و"موقع المفعول" لترميز بناء الجملة وفقًا لهذا النوع من المعلومات؛ كما يمكن أيضًا أن يعتمد هذا النوع من الترميز على التدرج البنائي للنص أو شرح العلاقة بين المتحدثين كمفردات للترميز.

وما نرغب في التأكيد عليه هنا هو علاقات التدرج البنائي للنص، وهي إحدى منهجيات الترميز على مستوى الخطاب التي تم تنفيذها حتى الآن بشكل عملي على ذخيرة ذات نطاق محدود. وتجدر الإشارة في هذا السياق

إلى أن عدد أنظمة الترميز على مستوى الخطاب قليلة إلى حدٍّ كبير حتى الآن. إن عملية ترميز الخطاب باتباع علاقات التدرج البنائي للنص ليست وثيقة الصلة بأساليب الترميز الأخرى كالترميز الصرفي والنحوي والدلالي. فالترميز الدلالي يهتم بإيضاح معاني الكلمات، أما ترميز الخطاب فيهتم بإيضاح الأوجه المتعددة للكفاءة اللغوية، ويُقصَد بذلك تلك الأمور المتعلقة بانتقال المعنى من جزء إلى آخر داخل النص. فإذا لم نتمكن من شرح المعنى على هذا المستوى، أصبح أمامنا عائق في فهم هذا الجزء من لغة البشر.

وقد سبق أن نوقش موضوع الترميز على مستوى الخطاب بشكل مكثف في أثناء مؤتمر 'DAAR96'، حيث إن الحاجة إلى تفسير الضمائر وما تشير إليه، وإزالة اللبس المتعلق بهذا الأمر من موضوعات البحث الرئيسة التي يهتم بها كل من علماء اللغة التقليديين وعلماء اللغة الحاسوبيين على حدٍّ سواء. وعلى الأخص فإن علماء اللغة الحاسوبيين بدءوا في الاهتمام بالذخائر اللغوية التي تُستَخدم في التدريب على هذا النوع من الترميز؛ التي تُستَخدم في اختبار البرمجيات. ففي العقود القليلة الماضية، أصبحت إشكاليات تحديد ما تشير إليه الضمائر إحدى القضايا الساخنة في مجالات الترجمة الآلية واستخراج المعلومات من النصوص. على سبيل المثال، ما تشير إليه ضمائر الغائب she،he،they،ti في أحد النصوص التي تتهيأ للمعالجة الآلية. وحتى يمكن التغلب على تلك المشكلة كان هناك وجهتان للنظر: الأولى ترى أنه لا غنى عن تغذية الآلية بالمعلومات اللغوية ومعلومات العالَم الحقيقي حتى تتمكن الآلة من تحديد ما تشير إليه تلك الضمائر، أما وجهة النظر الثانية

فترى أن الاعتماد على أسلوب الخبرة من الممكن أن يُوَصِّل إلى تلك النتيجة، وهذه المنهجية لا تَسْتَخدم معلومات العالم الحقيقي، بل تكتفي فقط باستخدام المسافة بين الضمير والكلمات السابقة له، بالإضافة إلى المعلومات الإحصائية عن الكلمات أو العلامات داخل النص في التوصل إلى ما تشير إليه الضمائر. إن الاستعانة بالمنهجية الإحصائية وذخيرة مُرَمَّزة مناسبة من الممكن أن يساعد في تجريب وسيلة آلية تعتمد فقط على النصوص في تمييز علامات الترميز الخطابي داخل النص.

قامت جامعة لانكاستر ببناء ذخيرة شجرية لإحالات الضمائر بتمويل من شركة آي بي إم. وكانت عبارة عن ذخيرة شجرية تحمل ترميزًا على مستوى أبنية الجمل، وعلى هذا الأساس تمت إضافة رموز خطاب تشير إلى علاقات التدرج البنائي داخل النص. وأثبتت تجارب الترميز أن الترميز على مستوى الخطاب يمكن تنفيذه بطريقة مُوَحَّدة ويتم الحصول على نتائج سليمة. وفيما يلي نقدم بعض أمثلة حقيقية توضح نتيجة الترميز على مستوى الخطاب الذي قامت به هذه الجامعة.

مثال ١:

٦) the married couple 6) said that<REF=6 they were happy with<REF= 6 lot.

مثال 2:

7) this week's winner 7) said <REF=7 he had rung (8 <REF 7 his wife 8) and <REF=7,8 they had spoken to< REF=7,8:2 each other.

حيث يتم وضع الكلمات التي سبق ذكرها بين قوسين ويسبق هذه الكلمات رقم مُفَهْرَس، ولا يوجد لهذا الرقم إلا مرجعية واحدة داخل النص؛ أما الضمير الذي تتوب عنه هذه الكلمات فتوضع قبله علامة الترميز "REF= الرقم المفهرس" أي أنه يُظْهِر الضمير المقابل لتلك الكلمات داخل النص.

الباب الرابع

علم الذخائر اللغوية

والدراسات اللغوية

تم الاتفاق على أن علم الذخائر اللغوية هو ذلك العلم الذي يعتمد على أسلوب جمع النصوص اللغوية الواقعية لدراسة مشكلات علم اللغة. فمن حيث المنهجية، نجد أن هذا الأسلوب ينتمي إلى المنهج التجريبي في البحث العلمي، الذي يختلف عن المنهج العقلي الذي اتبعه تشومسكي. ومن هنا، فقد اهتم الجميع اهتمامًا كبيرًا بتلك الدراسات اللغوية القائمة على ذخائر لغوية. وفي حقيقية الأمر، تميزت الدراسات اللغوية الصينية على مدى التاريخ دائمًا بالانطلاق من الوقائع اللغوية الحقيقية. إلا أن قصور متابعة الباحثين الصينيين للبنية المعرفية على مستوى العالم قد تسبب في تأخر الأبحاث الصينية القائمة على ذخائر لغوية مُمَيْكَنة في عمل دراسات لغوية لبعض الوقت عن الغرب.

الفصل الأول

منهجية الذخائر اللغوية في الدراسات اللغوية

يتناول هذا الباب بالتعريف أهم تطبيقات منهجية الذخائر اللغوية في الدراسات اللغوية. إن أحد أهم استخدامات منهجية الذخائر اللغوية في الدراسات اللغوية استخراج البيانات اللغوية التجريبية الأكثر شيوعًا وتقديمها للعاملين في مجال البحث اللغوي. ويبدأ هذا الباب بتعريف القارئ بالتطبيقات المختلفة لمنهجية الذخائر اللغوية في حقل الدراسات اللغوية، ويُختتم الباب بتقديم بعض الأمثلة الحقيقية التي حققت نتائج على أرض الواقع.

أولاً: الذخائر اللغوية وتطبيقاتها في الدراسات المتعلقة بعلم المفردات

إن تاريخ اعتماد مؤلفي المعاجم على البيانات اللغوية الواقعية واستخدامهم لها في مؤلفاتهم المُعْجَمية يسبق ظهور علم الذخائر اللغوية. على سبيل المثال سبق أن استخدم العالم صمويل جونسون (Samuel Johnson) الجمل الواردة في الأعمال الأدبية في تأليف معجمه. وفي القرن التاسع عشر، استخدم معجم أكسفورد للغة الإنجليزية (Oxford English Dictionary) بطاقات الاستشهاد (citation slips) لدراسة الاستخدامات المختلفة للكلمات وشرحها. وما زالت طريقة جمع الاستشهادات اللغوية من اللغة الواقعية مستمرة حتى الآن، إلا أن ظهور الذخائر اللغوية وما صاحبها من منهجيات قد غيَّر من أسلوب استقراء مؤلفي المعاجم واللغويين للحقائق اللغوية.

فالذخائر اللغوية في الوقت الراهن تعني أن مؤلفي المعاجم بإمكانهم الجلوس أمام إحدى شاشات الحواسيب الإلكترونية، وفي ما لا يزيد عن عدة ثوان يمكنهم استخراج الأمثلة الكاملة التي تُمَثِّل الاستخدام الحقيقي لكلمة أو تعبيرة لغوية في نصوص يتعدى حجمها مليون كلمة. وهذا لا يعني فقط أن إنتاج المعاجم وتطويرها يتم الآن بشكل أسرع مما سبق، بل يعني أيضًا أن تعريف المواد اللغوية يتم بشكل أدق وأكثر شمولية؛ وذلك لأن المواد اللغوية يتم استقراؤها من خلال ذخيرة لغوية حقيقية، وهذه الذخيرة تمثل تجمع هائل للعينات اللغوية يفوق كثيرًا ما كان عليه الوضع فيما مضى.

إن استخراج الأمثلة من الذخيرة اللغوية من شأنه أن يمثل تَجَمُّعًا لغويًّا مهمًّا يمكن استخدامه في تحليل أعمق لمعاني المفردات وتمييزها. على سبيل المثال: عمل ترتيب أبجدي للكلمات المصاحبة لكلمة ما من جهة اليمين؛ ومن ثَمَّ يمكن عمل استقراء للأمثلة الحقيقية عن جميع التصاحبات اللغوية التي تأتي مع هذه الكلمة في سياق لغوي معين. وبالإضافة إلى ذلك فإن الذخائر اللغوية التي يستخدمها مؤلفو المعاجم تحتوي معلومات تصنيفية حول مجموعة هائلة من النصوص اللغوية، مثل اسم مؤلف النص، وجنسه، وتاريخ النشر، والشكل اللغوي له، لدرجة أن تلك النصوص تكون مُرَمَّزَة: تحتوي ترميزًا للكلمات من حيث النوع والمعنى. وهذه المعلومات من شأنها أن تتيح إجراء تصنيف للمعلومات التي يتم استخراجها من الذخيرة، وهذا يساعد المعجميين في تحديد الاستخدامات المختلفة لكلمة ما في بعض المجالات والأشكال اللغوية المختلفة بصورة نموذجية.

من خلال البرمجيات التي تقوم بحساب درجة التصاحب بين الكلمات (سبق الإشارة إليها في الفصل الثالث من الباب الثالث) يمكن استخراج مجموعات الكلمات المتصاحبة، وهذا يعني أنه بالإمكان استقراء التعبيرات والتصاحبات اللغوية ومعالجتها بشكل أكثر مصداقية مما سبق. فالوحدات اللغوية شائعة الاستخدام (phraseological) من الممكن أن تُكَوِّن مصطلحا تقنيًا، وفي أحيان أخرى قد تكون مثلاً شعبيًا، أما التصاحبات اللغوية فتعتبر خيطًا مهمًا في التعرف على معنى الكلمة موضع البحث[68]. إن تمييز هذه التصاحبات داخل النصوص يعني إمكانية التعامل معها بشكل جيد كما يحدث مع الكلمات المفردة؛ حيث يتم معالجتها في المعاجم أو في بنوك المصطلحات الآلية، التي تُستَخدَم لإمداد المترجمين والاستعانة بها عند الحاجة.

وفي الوقت ذاته، فإن صناعة المعاجم القائمة على ذخائر لغوية تساعد المعجميين في استخراج التعاريف اللغوية من الذخائر اللغوية؛ كأن يتم استخدام التصاحبات شائعة الاستخدام في ربط المعاني المتعلقة بالكلمات معًا، وهذا يساعد المعجميين في تقسيم فهارس الاستخدام للكلمات إلى تصنيفات تمثل الاستخدامات المختلفة للكلمة؛ ومن ثَمَّ يمكن تقديم معلومات عن معدلات تواتر المعاني المختلفة للكلمات من حيث معدلات التكرار.

ثانيًا: الذخائر اللغوية والنحو

تُعتَبر الدراسات النحوية (أو الدراسات المتعلقة بتركيب الجمل) مثلها كمثل الدراسات المتعلقة بعلم المفردات، إحدى الأمثلة الواقعية التي تشير إلى

اعتماد الدراسات اللغوية على الذخائر اللغوية. وتظهر أهمية الذخائر اللغوية في الدراسات المتعلقة بتركيب الجمل في النقاط التالية:

(١) تُعتَبر الذخائر اللغوية بمثابة تمثيل للغة بأكملها.

(٢) باعتبار أن البيانات اللغوية التي تُقَدِّمها الذخائر اللغوية بيانات تجريبية، فإن الحقائق اللغوية التي تُقَدِّمها يمكن استقراؤها بالأساليب الإحصائية.

قبل ثمانينيات القرن العشرين، كانت الدراسات اللغوية التجريبية تضطر إلى الاعتماد بصورة أساسية على أساليب التحليل الثابتة. وكان هذا النوع من الدراسات يُقَدِّم وصفًا دقيقًا للمنظومة النحوية للّغة؛ ولكن النتائج كان من الصعب أن تَرصُد معدلات التكرار الأعلى والأقل بشكل موضوعي. ومع ظهور الذخائر اللغوية المُرَمَّزة على مستوى تركيب الجملة، والتطور المستمر لأدوات البحث داخل الذخائر اللغوية، أصبح من السهل إجراء التحليل الكمي (**quantative analysis**) للظواهر النحوية بشكل أكبر مما سبق. إن التحليل الكمي للظواهر النحوية على أقل تقدير يقدم للباحثين أفضل نماذج الاستخدام النحوي لتلك الظواهر، بالإضافة إلى كل درجات التحول التي تحدث وما إلى ذلك من معلومات. وهذه المعلومات لا تفيد فقط في فهم القواعد النحوية للغة ما، بل تفيد أيضًا في دراسة أوجه الاختلاف بين اللغات بعضها البعض، وفي مجال تعليم اللغات.

إن معظم الدراسات النحوية محدودة النطاق التي اعتمدت على الذخائر اللغوية كانت تشمل تحليلاً كميًّا للبيانات؛ على سبيل المثال الدراسة التي قام بها العالم شميت (Schmidt:١٩٩٣)[69] عن الربط بين العبارات توصلت إلى العديد من المعلومات الكمية عن هذه الظاهرة في ذخيرة لــوب **LOB** للغــة الإنجليزية. وتفيد منهجية الذخائر اللغوية في إحصاء معدلات تكرار أنمــاط الجمل كافة.

منذ خمسينيات القرن العشرين، انقسم علماء اللغة إلى قسمين كبيــرين: القسم الأول تبنى في دراسة الظواهر اللغوية المنهج العقلــي (**rationalism**)، والقسم الثاني استخدم الأسلوب الوصفي التجريبي في دراسة الحقائق اللغويــة، وأولى اهتمامًا كبيرًا بالإحصاء الكمي للظواهر اللغوية في الذخائر اللغوية. إلا أن هذين الفريقين لا ينكر أحدهما الآخر كما يعتقد البعض. ففي الحقيقة، هناك بعض الباحثين ممن ينتمون إلى الفريق الأول استخدموا الذخائر اللغويــة فــي قياس النظريات النحوية التي تنتمي إلى المنهج العقلي. ولم يلجئوا إلى استخدام الذخائر اللغوية في عمل التوصيف اللغوي البحت أو توليد النظريات النحوية.

ففي جامعة نيجمن (**Nijmen**) الأميركية تم الجمع بين كل من المنهج العقلي والمنهج التجريبي في دراسة النحو لبناء نظريــة نحويــة صــورية (**formal Grammar**) تخدم المنهج العقلي. وبعد ذلك تم اختبار تلك النظرية النحوية على الواقع اللغوي المُخزَّن في الذخائر اللغوية الإلكترونية. وكــان ذلك عن طريق البدء بالاطلاع على آراء جمهور النحاة فــي شــرح تلــك القواعد واستخدام آرائهم في تصميم نموذج للنحو الشكلي، وبعد ذلك تم تغذية

برمجية التحليل النحوي الحاسوبية بهذا النموذج، واستخدامها في تحليل الذخيرة اللغوية؛ للوقوف على مدى قدرة البرمجية مستعينة بهذا النموذج في تحليل البيانات اللغوية الموجودة بالذخيرة. وعلى الأساس التجريبي لنموذج التحليل النحوي هذا تم تصويب هذا النموذج من خلال الأجزاء التي لم يتم تحليلها أو التي حُلِّلَت بصورة خاطئة.

وهناك تفكير آخر للاستفادة من الذخائر اللغوية في عمل التحليل النحوي؛ وذلك عن طريق الاستعانة بذخيرة مُرَمَّزَة على مستوى بنية الجمل، ومن خلال البيانات الإحصائية يتم تصنيف القواعد النحوية الواقعية الموجودة بها. ويمكن الاطلاع على مثال عن هذه المنهجية في الفصل الثاني من الباب الرابع.

ثالثًا: الذخائر اللغوية وعلم الدلالة

تعرفنا من خلال الفصول والأبواب السابقة أنه يمكن الاعتماد على ذخيرة لغوية في استقراء الحالات التي تظهر فيها كلمة ما للتعرف على معناها. وهذه المنهجية تُستَخدم بصورة أساسية في مجال علم المعجمية. ولكن بصورة عامة يمكننا القول: إن الذخائر اللغوية تلعب دورًا مهمًّا في خدمة علم الدلالة؛ حيث يبرز دورها في إمداد علم الدلالة بشروح موضوعية تعتمد على أسلوب ديناميكي يتغير حسب طبيعة التغيرات اللغوية. ويتمثل أول دور مهم للذخائر اللغوية في علم الدلالة في إمكانية حصر المعاني الإضافية للكلمات بشكل موضوعي وفقًا للواقع اللغوي. وقد سبق أن أشار العالم اللغوي منِت (Mindt:١٩٩١)[70] إلى أن أقسام معاني الكلمات في علم الدلالة والمعاني الأكثر تواترًا في البناء اللغوي يتم تقريرها ووصفها وفقًا

لإدراك علماء اللغة، وهذه هي المنهجية العقلية في التعامل مع اللغة. ولكــن الحقيقة أن تمييز المعنى ينبغي أن ينطلق من اســتقراء الــسياقات والأبنيــة النحوية للجمل وعلم الصرف وخصائص علم العروض، واستخدام الــذخائر اللغوية من شأنه أن يقدم أمثلة موضوعية على تلك الفروق الدلالية.

أما الاستخدام الثاني المهم للذخائر اللغوية في علم الدلالة هو تقديم بنية أكثر تماسكًا فيما يتعلق بالتصنيف الضبابي (yfuzzy categor) وفكرة التحول التدريجي في المعنى. ففي علم اللغة النظري يُنظَر عــادة إلــى التــصنيف باعتباره تقسيمًا ثابتًا لا يمكن تغييره. وهذا يعني أنه عند التعامل مع كلمــة معينة، إما النظر إليها على أنها تنتمي إلى نطاق تصنيفي معين، أو لا تنتمي إليه. ولكن الدراسات التي أجريت في فرع علم النفس التــصنيفي تــرى أن فئات الإدراك لا تتصف بالثبات، بالإضافة إلى أن الحدود فيما بينها ضبابية إلى حد بعيد. ولذلك، فالمشكلة لا تكمن في إقرار ما إذا كانت كلمة تنتمي إلى تصنيف معين أم لا؟ ولكن في احتمال ظهور هذه الكلمة ضمن تصنيف مــا مقارنة مع احتمال ظهورها ضمن تصنيف آخر. ولا يمكن الحصول علــى هذه المعلومات الاحتمالية بأي حال من الأحوال إلا باستخدام الذخائر اللغوية التي تمثل الواقع اللغوي الحقيقي.

رابعًا: الذخائر اللغوية في مجال علم اللغة التداولي وتحليل الخطاب

الدراسات التي تهتم بتحليل الذخائر اللغوية على أســاس علــم اللغــة التداولي وتحليل الخطاب حتى الآن قليلة إلى حدٍّ بعيد. والسبب الرئيس فــي ذلك يرجع إلى أن مصدر التحليل على المستويين التداولي والخطابي يُــستَمَد

من العلاقة السياقية بين الجمل بعضها البعض. وعلم اللغة التداولي دائمًا ما يُطلَق عليه "علم المعنى السياقي"، إلا أن النصوص المُمَثَّلة في الذخائر اللغوية تفقد جزءًا كبيرًا من السياق الخاص بها. والسبب في ذلك يرجع إلى أن الذخائر اللغوية تتجه إلى جمع عينات لغوية صغيرة الحجم وليس نصوصًا كاملة. هذا بالإضافة إلى أن العينات النصية التي تُجْمَع يتم حذف السياقات الاجتماعية والنصية الخاصة بها.

وقد تركز جزء كبير من الدراسات المتعلقة بعلم اللغة التداولي والمجالات المتعلقة به خارج الصين على اللغة الشفهية. وتُعْتَبَر ذخيرة لندن-لوند **(Lund-London)** الذخيرة الوحيدة التي تضم نصوصًا حوارية. ولذلك فإن غالبية تلك الدراسات تم إنجازها على هذه الذخيرة. وكانت أهم الإسهامات التي قدمتها تلك الدراسات هي فهم الكيفية التي يتم بها الحوار اللغوي، وعلى الأخص العلاقة بين الكلمات والتعبيرات اللغوية والجمل المكوِّنة للحوار. وقد قدمت العالمة ستنستورم **(Stenstorm:1987)[71]** من خلال الدراسة التي قامت بها على ذخيرة لغوية تفسيرًا كميًّا من وجهة نظر علم التصنيف حول عينات من الحوارات التي يبادر الشخص بإصدارها وأخرى يكون الشخص مشاركًا من خلالها في حوار طبيعي. على سبيل المثال، في أثناء استقرائها للرسائل اللغوية التي تحمل كلمة "right"، وجدت أن تعبيرة "all right" دائمًا ما تأتي بين الحدود الفاصلة بين مرحلتين من مراحل الحوار؛ وأن "that's right" غالبًا ما تُستَخدم في رسالة لغوية تحمل معنى التوكيد، أما "it's right"، و"that's right" فتحملان معنى التجاوب

الحواري بالإضافة إلى التعبير عن معنى التواضع. وهذا الأسلوب الكمي في دراسة أساليب الحوار عَمَّق من مستوى المعرفة البشرية تجاه عدد من التصرفات اللغوية. وذلك لأن تلك النتائج قدمت شروحًا تخصصية بما في ذلك الخيارات التي تكون أمام المتحدث من خلال وجهة النظر التداولية وأي من تلك الخيارات أكثر قياسية وأيها قليل التداول.

خامسًا: الذخائر اللغوية وتعليم اللغات

عند تأليف المقررات الخاصة بتعليم اللغة دائمًا ما يكون هناك اتجاهان مختلفان، أحدهما يمثل المنهجية التجريبية وآخر يمثل المنهجية العقلية. فمن ناحية نجد أن العديد من المقررات الدراسية تَسْتَخْدِم الجمل المُؤَلَّفة؛ حيث تعتمد تلك الجمل على إدارك من يقوم بتأليفها، وفي الوقت نفسه تمثل مواد لغوية سبق استخدامها. ومن ناحية أخرى، نجد أن الأمثلة التي تقدمها المعاجم والمؤلفات المساعدة في تعليم اللغة؛ التي تم بناؤها على ذخيرة كولينز كوبويلد (Collins-COBUILD) تم جمعها على العكس من ذلك باستخدام المنهج التجريبي. وقد اعتمدت الأمثلة التي تم تقديمها من خلال تلك التجربة على مصادر حية من الذخيرة اللغوية أو من مصادر لغوية أخرى.

إن الذخائر اللغوية تُعْتَبَر مصادر مهمة للأمثلة في عملية تعليم اللغة؛ وذلك لأن الدارسين في المرحلة المبكرة من دراسة اللغة يكونون في حاجة إلى الجمل والمفردات الحقيقية؛ حيث إنهم سيتعرضون لهذه المفردات في أثناء اطلاعهم على النصوص التحريرية الحقيقية لتلك اللغة وسيحتاجون إليها

في أثناء عملية التواصل اللغوي. وهذه البيانات التجريبية لها الأهمية نفسها في تعليم اللغة للناطقين بها أو للأجانب؛ حيث تتجاوز فائدتها نطاق المعنى البسيط الذي يدور حول الاستخدام الواقعي للغة. وهناك عدد غير قليل من العلماء اعتمدوا على بيانات الذخائر اللغوية في توجيه النقد لمقررات تعليم اللغة الموجودة حاليًا.

وكانت الطريقة التي اتبعوها في ذلك متشابهة إلى حدٍّ بعيد؛ حيث استخدموا ذخيرة قياسية للغة الإنجليزية مثل ذخيرة لوب **LOB** وذخيرة لندن- لوند **(Lund–London)** في عمل تحليل لمحتوى المقررات الدراسية أو قوائم الكلمات المستخدمة فيها. ثم قارنوا مجموعتي النتائج التي توصلوا إليها. فتوصلت معظم الدراسات إلى أن هناك اختلافات جوهرية بين محتوى تلك المقررات الدراسية والاستخدام الحقيقي الذي يمارسه به أهل تلك اللغة. كما أن هناك بعض المقررات الدراسية أهملت جوانب متعددة لبعض أوجه الاستخدام اللغوي المهمة. لدرجة أنه في بعض الأحيان تم التضحية بأكثر الأساليب شيوعًا في اللغة، في مقابل إظهار أساليب أخرى ليس لها مثل هذه الدرجة من الشيوع. وكانت النتيجة التي استخلصها كل من العالمين الشهيرين مِنت **(Mindt)** وكينيدي **(Kennedy)** وغيرهما أن المقررات التي لا تقوم على أساس من البيانات اللغوية التجريبية من المؤكد أنها ستسير في الطريق الخطأ. فمن الضروري اللجوء إلى استخدام الذخائر اللغوية في توجيه عملية تأليف المقررات الدراسية؛ ومن ثَمَّ إعطاء مزيد من الاهتمام للحقائق اللغوية الأكثر شيوعًا من حيث الاستخدام.

وهناك منهجية خاصة في تعليم اللغات الأجنبية يُطلَق عليها اسم "تعليم اللغة للأغراض الخاصة" وتشير تلك الطريقة إلى تعليم اللغة لأغراض تختص ببعض مجالات الاستخدام اللغوي، وهذا ما يتعارف عليه الجميع باسم تعليم اللغات الأجنبية لأغراض تخصصية؛ مثل تعليم اللغة الإنجليزية لطلاب الطب. إن بناء ذخائر لغوية متعددة المجالات من الممكن أن يساهم في تعليم اللغة لتلك الأغراض الخاصة كل في تخصصه. وقد سبق أن قامت جامعة العلوم والتكنولوجيا بهونج كونج ببناء ذخيرة للغة الإنجليزية مكوَّنَة من مليون كلمة. وكانت العينات المستخدمة منتقاة من المقررات الدراسية التي يستخدمها طلاب تخصص الحاسب الآلي. وبإمكان تلك الذخيرة أن تُقَدِّم للطلاب العديد من المواد في المجالات التخصصية؛ بما في ذلك قوائم المفردات والبيانات الكمية عن طرق الاستخدام اللغوي. وتعمل هذه المواد على الاحتياجات الخاصة للطلاب في هذا المجال التخصصي، وهذا بالطبع أفضل من تلك التي يتم استخراجها من ذخائر لغوية في المجالات العامة.

الفصل الثاني

إحصاء أنماط الجمل في اللغة الصينية والدراسات القائمة عليه

تم ذلك في إطار مشروع رسالة دكتوراه تم تمويلها من صندوق اللجنة القومية للتعليم بإشراف الأستاذ الدكتور جاو شو خوا الأستاذ بجامعة اللغات والثقافة ببكين، وقد نوقشت الرسالة واعتُمِدَت من قِبَل مجموعة من الخبراء في بكين في العاشر من شهر إبريل عام ١٩٩٥[72].

١- كانت الأهداف الرئيسة من دراسة إحصاء أنماط الجمل في اللغة الصينية ما يلي:

(أ) إجراء تقطيع على مستوى الجمل لمواد لغوية مُكَوَّنَة من نصوص المقررات الدراسية الصينية في المرحلتين الابتدائية والإعدادية بالإضافة إلى المقررات الدراسية لتعليم اللغة الصينية للأجانب بجامعة اللغات والثقافة، وقد وصلت النصوص التي تم تقطيعها إلى أربعة ملايين رمز صيني، ومن نتائج التقطيع تم عمل منظومة لتصنيفات الجمل في اللغة الصينية الحديثة. وبناءً على ذلك، تم إحصاء معدل تكرار كل نوع من الجمل داخل الذخيرة؛ ومن ثَمَّ

أمكن تقديم بيانات علمية يمكن الاعتماد عليها في الدراسات متعددة الجوانب الخاصة بتعليم اللغة الصينية للأجانب، وتصميم المقررات، ووضع معايير الاختبارات اللغوية، بالإضافة إلى الدراسات التقابلية بين اللغة الصينية وغيرها من اللغات.

(ب) تقديم ذخيرة لغوية مؤهّلة تحتوي البيانات اللغويـة الـضرورية لاستخدام علماء اللغة والعاملين في مجال تعليم اللغة الصينية.

(ج) تقديم مصادر أساسية تُستخدم في مجالات التمييز الآلي لكل مـن حدود التعبيرات اللغوية، وفصل عناصر الجملة آليًّا، بالإضافة إلى تحليل التركيب النحوي للجمل الصينية آليًّا وغير ذلك من الدراسات المتخصصة في مجال المعالجة الآلية للمعلومات باللغة الصينية.

٢- وقد حقق هذا المشروع البحثي النتائج التالية:

(أ) عمل تقطيع على مستوى الجمل لمحتويات ذخيرة من أربعة ملايين كلمة، وتخزين الناتج في صورة رقمية، بالإضافة إلى تحويلها إلى بطاقات ورقية تصل إلى مائتي ألف بطاقة للأمثلـة مـن الجمـل اللغوية مُفَهرَسة حسب نوع الجملة، وموضَّح عليهـا المعلومـات التوثيقية الخاصة بها.

(ب) عمل تصنيف مشفوعٍ بالإحصاءات عن أنماط الجمـل المكوّنـة للمادة اللغوية التي تخص نصوص اللغة والأدب المـستخدمة فـي المرحلة الابتدائية؛ التي تصل إلى مليونين وثمانمائة رمز وتقسيمها

إلى ثلاثة مستويات مبتدئ ومتوسط ومتقدم. وفي الوقت نفسه عمل تحليل لتركيب الجمل البسيطة التي تم استخراجها وتخزين النتائج في شكل رقمي، وبذلك تم تكوين بنك عن أنماط الجملة البسيطة في اللغة الصينية الحديثة. وقد وصل عدد الجمل البسيطة التي تم تخزينها إلى ١٦٢٩٧ جملة. وبإمكان بنك أنماط الجمل المشار إليه أن يُقَدِّم معلومات عن الخصائص التركيبية لكل نوع من أنواع الجمل الصينية والأمثلة المقابلة لهذه الأنواع، بالإضافة إلى المعلومات التالية:

١- معدلات استخدام الأنواع المختلفة للكلمات الصينية في الأماكن التي تحتلها عناصر الجملة الصينية الستة وهي المسند إليه، والمسند، والنعت، والحال، ومُكَمِّل المعنى، بالإضافة إلى توجهات المعنى التي يشير إليها كل من النعت والحال والمُكَمِّل في الجملة.

٢- أنواع الكلمات التي تظهر بين حدود كل من المسند والمفعول به في الجملة ومعدلات ظهورها.

٣- ترتيب العناصر في كل من الحال المُوَسَّع والنعت المُوَسَّع.

٤- الجزء الحاوي للرمز "的" في النعت المُوَسَّع، الجزء الحاوي للرمز "地" في الحال المُوَسَّع.

٥- الأجزاء التي تحتوي الأدوات المساعدة المُبَيِّنة لحالة الفعل "了，着，过，来着" وما إلى ذلك.

٦- الأمثلة التي تحتوي فعل ومفعول به وما إلى ذلك من معلومات.

(ج) عمل تحليل جزئي لتراكيب الجمل التي تحتويها نصوص مادة المطالعات الصينية بجامعة اللغات والثقافة؛ التي يصل حجمها إلى ٣٤٠،٠٠٠ كلمة، بالإضافة إلى انتقاء الجمل الصعبة من تلك النصوص وعمل تحليل لجزء منها على مستويات ثلاثة: مستوى تركيب الجمل، ومستوى الدلالة، ومستوى التداولية. وعمل إحصاء عن معدلات استخدام بعض أنماط الجمل وبعض الأفعال ذات الطبيعة الخاصة.

(د) عمل تصنيف للجمل التي ظهرت في نصوص اللغة والأدب للمرحلة الابتدائية، نتج عن ذلك التصنيف (المسرد الإحصائي لمعدلات تكرار أنماط الجمل الصينية)، و(مسرد أنماط الجمل الشائعة في اللغة الصينية الحديثة). وهذين المسردين لهما قيمة مرجعية عالية في مجالات الدراسات التقابلية بين أنماط الجمل في اللغة الصينية واللغات الأخرى، ووضع معايير اختبارات تحديد مستوى إجادة اللغة الصينية، بالإضافة إلى تأليف مقررات اللغة الصينية للأجانب. على سبيل المثال، فيما مضى كانت مقررات تعليم اللغة الصينية للأجانب في أثناء شرح المُكمِّل المبين للحالة الشعورية (أي المكمل الذي يحتوي الرمز "得"، ويُطلَق عليه أيضًا المُكمِّل المبين للدرجة) وعادة ما كانت تلك المقررات تشرح هذا النوع من المكملات من خلال مثال توضيحي عن طريقة تركيب

هذا المكمل: "مسند إليه+ فعل+ مفعول به+ تكرار للفعل+ 得 + المكمل" مثل "他写字写得快" بمعنى "هو يكتب الرموز الصينية بسرعة" وفي الحقيقة أن هذا التركيب قليل التواتر في الاستخدام اللغوي الواقعي، ففي نصوص اللغة والأدب للمرحلة الابتدائية التي تصل إلى ٢٨٠،٠٠٠ رمز لم يظهر هذا التركيب ولو مرة واحدة. ولذلك يرى فريق البحث أن الأمر يحتاج إلى إعادة نظر في مدى ضرورة شرح هذا النوع من الجمل في المستوى المبتدئ من مستويات تعليم اللغة الصينية للأجانب.

٣- وقد توصل فريق البحث من خلال هذه الدراسة إلى القواعد التالية في تعريف الجملة:

(أ) ضرورة وجودها في إطار بنائي متكامل.

(ب) تعبيرها عن معنى متكامل.

(ج) تمتُّعها بتنغيم يجعلها مستقلة بذاتها.

٤- وقد قابل فريق البحث عددًا كبيرًا من الجمل المركبة في أثناء عملية تقطيع نصوص الذخيرة إلى جمل، وقد عالجوا تلك الجمل كما يلي:

(أ) إذا كانت الجمل الفرعية المُكوِّنة للجملة المركبة يمكن أن تستقل بذاتها، يتم التعامل معها على أنها جملة بسيطة (مع إغفال أدوات الربط).

(ب) إذا كانت إحدى الجمل الفرعية يمكن أن تستقل في صورة جملة كاملة، يتم التعامل معها على أنها جملة بسيطة، والتعامل مع الجزء الباقي من الجملة على أنه جملة ناقصة.

(ج) إذا كانت الجمل الفرعية المُكَوِّنة للجملة المركبة لا يمكن أن تستقل بذاتها، يتم التعامل مع تلك الجملة على أنها جملة مركبة. بالإضافة إلى أن الجمل الناقصة والجمل المختصرة المُكَوِّنة للجمل المركبة لم يتم احتسابها في نتيجة إحصاء أنماط الجمل الصينية.

في المادة اللغوية الممثِّلة لنصوص اللغة والأدب في المرحلة الابتدائية تم تقطيع عدد ١٤٣٨٧ جملة، وعدد ١٣٢ جملة مختصرة، وعدد ٥٧٨ جملة ناقصة (مثل: "傍晚会到家" بمعنى "يعود إلى المنزل مساءً"، و"张开嘴正要吃" بمعنى "ما إن فتح فمه وبدأ في الأكل"، و"他抬头一看" بمعنى "ما إن رفع رأسه ليرى")، وعدد ١٢٠٠ جملة مركبة (مثل "有过一道湾，又过了一道湾"، بمعنى "خرج من متاهة، ليدخل في متاهة أخرى"، أو "他一会儿弯弯腰，一会儿压压腿 "، بمعنى " يثني وسطه تارة، ويُدلِّكُ ساقه تارة أخرى" أو "只要命令一下他们，就按动扳机" بمعنى "ما إن تصدر إليهم الأوامر، حتى يشغلوا لوحة التحكم"، وما إلى ذلك).

انحصرت الدراسة التي قام بها فريق البحث عن أنماط الجمل على دراسة أنماط تراكيب الجمل الصينية. وكانت القرائن التي يتم الاعتماد عليها

في التمييز بين أنماط الجمل المختلفة تتمثل في طبيعة الأجزاء المُكَوِّنة للجملة وطريقة بناء الجملة، بما في ذلك ترتيب الكلمات داخل الجملة وعدد العناصر المكونة للجملة، بالإضافة إلى أنواع الكلمات المُمَثِّلة لتلك العناصر، وما إلى ذلك.

٥- وقد تعمق فريق البحث في الدراسة وقَدَّمَ دراستين تكميليتين عن خصائص تركيب الجمل في اللغة الصينية:

(أ) على الرغم من أن تقسيم أنماط الجمل قد تم وفقًا لمجموعة من الأدلة عن الخصائص التركيبية للجمل، فإن فريق البحث لم يستبعد دراسة العلاقة الدلالية بين أجزاء الجملة عند الضرورة. على سبيل المثال في الجملة الإسنادية، على الرغم من أن تركيب جميع الجمل يكون على الصيغة "مسند إليه كبير + مسند إليه صغير + مسند صغير"، فإن العلاقة الدلالية بين العناصر المُكَوِّنة لتلك الجملة غير موحدة، ومن الممكن تصنيف هذه الجملة إلى خمسة أنماط فرعية مختلفة. على سبيل المثال إذا كان المسند الصغير عبارة عن فعل أو تعبيرة فعلية مكوَّنة من فعل ومفعول، تكون العلاقة الدلالية بين العناصر الثلاثة لهذه الجملة كما يلي:

• ٥SP عندما يكون المسند إليه الكبير واقع عليه الفعل.

• ٦SP عندما يكون المسند إليه الصغير واقع عليه الفعل.

• ٧SP عندما يكون المسند إليه الصغير جزءًا من المسند إليه الكبير.

• ٨SP عندما يكون المسند إليه الكبير كلمة معبرة عن المكان.

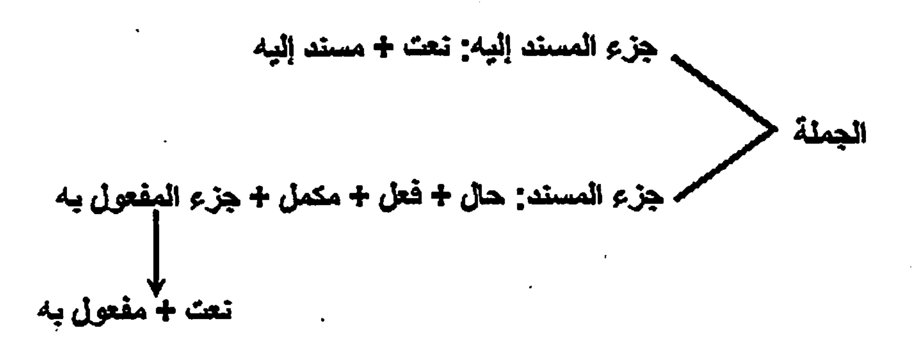
SP٩ • عندما يكون المسند إليه الصغير مشيرًا إلى المسند إليه الكبير.

(ب) حسب ما هو معروف لدى الجميع لا يُعْتَبَر النعت جـزءًا مـن أجزاء الجملة؛ نظرًا إلى عدم إمكانية تأثيره في البنيـة الأساسـية للجملة. ولكن في أثناء التطبيقات الحقيقية اكتشف فريق البحـث أن هناك بعض الجمل لا يستقيم لها المعنى إذا حُذِفَ منها النعت مثل: "叶公见了龙" بمعنى "رأى يه جونغ تِنْينًا" ولكن إذا أُضيف النعت تختلف الجملة تمامًا فتـصبح: "叶公看见了一条真龙" بمعنـى "رأى يه جونغ تِنْينًا حقيقيًّـا"، مثـال ذلـك أيـضًا جُمَـل مثـل "卖完了好大一叠报" بمعنى: "باع كومة كبيرة مـن الجرائـد"، و"受了一场虚惊" بمعنى "تعرض لخوف وهمي"، وما إلى ذلـك. ولذلك فإن النعت يشبه الحال في كونه ذا أهمية في تمييـز أنمـاط الجمل. ولذلك فإن تركيب هذا النمط من الجمل ينبغي أن يوضـع بالصيغة التالية: "مسند إليه+ فعل+ الأداة 了+ نعت+مفعـول بـه" ومثال آخر يتمثل في الجمل التي تحتوي الرمز "有" المُعَبِّر عـن معنى الملكية أو الوجود مثـل "那姑娘有一双漂亮的大眼睛" بمعنى "تلك الفتاة لها عينان واسعتان جميلتان"، ويكون تركيب هذه الجملة كما يلي: "مسند إليه + الرمز 有 + نعت (مكون من عـدد وكلمة كمية وكلمة وصفية) + مفعول به"، ويرجع ذلك إلى أنه في هذا النوع من الجمل لا يمكن الاستغناء عن النعت الـذي يـسبق الكلمة المحورية في المفعول به. فنحن لا يمكننا أن نكتفي بالقول:

"那姑娘有眼睛" بمعنى "تلك الفتاة لها عينان". وفي الحقيقة أن معنى هذه الجملة لا يعبر عن الملكية بقدر ما يؤكد معنى الوصف في الجملة مثل "那姑娘的眼睛又大又漂亮" بمعنى "إن عيني تلك الفتاة واسعتان وجميلتان".

وقد جمع فريق البحث بين منهجية التحليل حسب عناصر الجملة، ومنهجية التدرج في التحليل. ففي أثناء تقسيم عناصر الجملة استعانوا بنمط التركيب " نعت + مسند إليه + حال + فعل + مكمل + نعت + مفعول به"، مع اعتبار أن هناك علاقة تدرج بين العناصر وبعضها البعض، وأنها لا تقع في مستوى واحد. وعلاقة التدرج هذه تظهر خلال عملية تحليل الجملة.

٦- فالجملة يتم تقسيمها في البداية إلى جزأين كبيرين هما المسند إليه والمسند. يتكون جزء المسند إليه من "نعت + مسند إليه"، أما جزء المسند فيحتوي "حال + فعل + مكمل + مفعول به"، هذا بالإضافة إلى أن جزء المفعول به من الممكن أن يكون على النمط " نعت + مفعول به"، كما يظهر من الشكل التالي:



(أ) من الممكن أن تدخل التعبيرات اللغوية ضمن بنية الجمل كوحدة متكاملة، وتُكَوّن بذلك عنصرًا من عناصر الجملة. إلا أن التركيب الداخلي لهذه التعبيرات اللغوية ينبغي تحليله خارج إطار تحليل الجملة؛ وذلك لأن البنية التركيبية لتلك التعبيرات اللغوية لا تقع في المستوى البنائي نفسه الذي يحكم عناصر الجملة الستة (المسند إليه، والمسند، والمفعول به، والنعت، والحال، والمُكَمِّل).

(ب) جميع تراكيب التعبيرات اللغوية المكَوَّنة من نعت ومنعوت (يكون نمط التركيب داخل الجملة على هيئة " نعت + مسند إليه"، أو "حال + فعل + مكمل + مفعول به") يتحدد التدرج البنائي لها انطلاقًا من الكلمة المحورية في التعبيرة اللغوية؛ أي من اليمين إلى اليسار؛ أي أن التوسع في المعنى يحدث من الوحدة الصغرى في اتجاه الوحدة الكبرى. وجميع التعبيرات اللغوية التي تتكون من فعل ومكمل، أو فعل ومفعول، أو فعل ومفعول ومكمل، أو فعل ومكمل ومفعول، يتحدد التدرج البنائي لها انطلاقًا من الكلمة المحورية؛ أي من اليسار إلى اليمين، ويحدث التوسع في المعنى بدءًا من الوحدة الصغرى وفي اتجاه الوحدة الكبرى. إن الالتزام بهذه المبادئ من شأنه أن يُقَدِّم عرضًا شاملاً لمراحل تحليل الجملة الصينية.

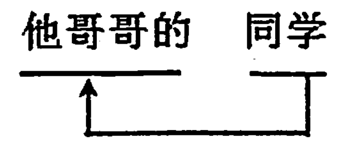
وفيما يلي نشرح مراحل تحليل الجملة الصينية من خلال مثال على جملة ذات مسند فعلي:

المثـــال "他哥哥的同学昨天在书店买到一本新出版的好词典"
بمعنى " اشترى صديق أخيه الأكبر أمس من المكتبة إصدارًا حديثًا لمعجم جيد".

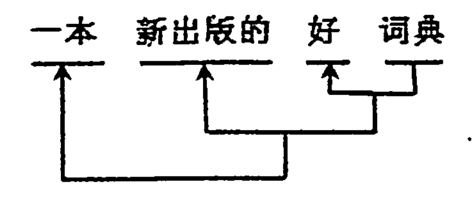
الخطوة الأولى: البدء بتحديد الكلمة المحورية في المسند، وهي الفعل "买" بمعنى " يشتري".

الخطوة الثانية: طبقًا للحدود الفاصلة بين جزأي المفعول به والمسند، يتم البحث عن عبارتي الصفة والموصوف الواقعتين في بداية الجملة ونهايتها، ثم الانطلاق من الكلمة المحورية وتمييز التركيب من اليمين إلى اليسار، ويتم التوسع وفقًا لتدرج البناء. على سبيل المثال:

تعبيرة "他哥哥" في "他哥哥的同学" هي تعبيرة من صفة وموصوف كما يوضح الرسم التالي:

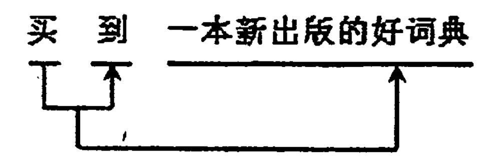


تعبيرة "新出版" في "一本新出版的好词典" هي تعبيرة حالية كما يوضح الرسم التالي:

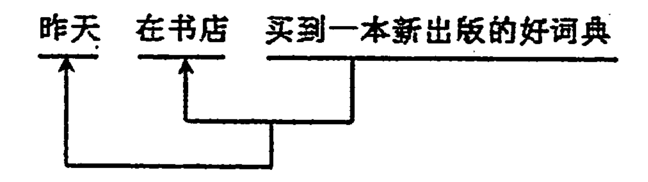


وإلى هذا المستوى يكون المسند إليه والمفعول به قد تم تحليلهما.

الخطوة الثالثة: يتم الانطلاق من الفعل باعتباره الكلمة المحورية في المسند، وتمييز التركيب من اليسار إلى اليمين، مثل:

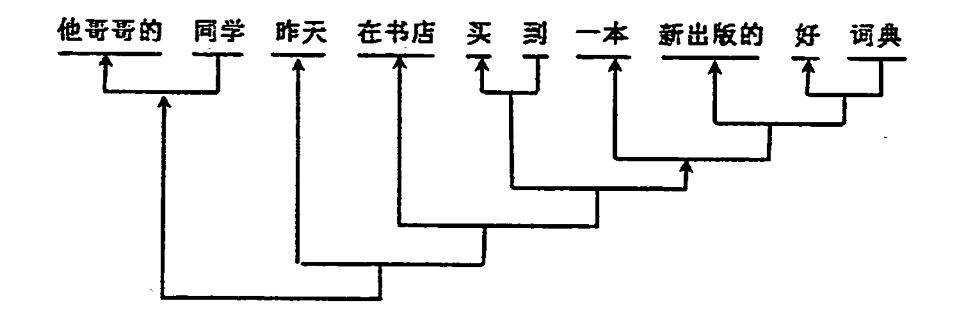


الخطوة الرابعة: الانطلاق من تعبيرة الفعل والمكمل والمفعول، وتمييز التركيب من اليمين إلى اليسار. مثل:



وبذلك يكون الجزء الخاص بالمسند قد تم تحليله.

الخطوة الخامسة: الجمع بين جزأي المسند إليه والمسند؛ أي تكوين الجملة كاملة؛ مثل:



الفصل الثالث

الأبحاث المتعلقة بعلم المعجمية

علم المعجمية هو أحد أفرع علم المفردات التطبيقي، ويهتم بدراسة مبادئ تأليف المعاجم والتطبيقات المتعلقة بهذا المجال، بمعنى إجراء التأليف المعجمي من خلال جمع الوحدات المعجمية (**lexical item**) ومقارنتها وشرحها وتصنيفها. إن الموضوعات الرئيسة التي يهتم بها علم المعجمية هي معاني الكلمات واستخداماتها. ويختلف علم تأليف المعاجم عن غيره من المجالات المختلفة لعلم اللغة في الغرب في أن هذا العلم اعتمد منذ نشأته وعبر مراحل تاريخه المختلفة على المنهج التجريبي القائم على الذخائر اللغوية. ففي عام **1755** اعتمد العالم جونسون (**Johnson**) على ذخيرة لغوية في جمع الاستخدامات الحقيقية للكلمات؛ حيث جعل من تلك الأمثلة مادة لتأليف معجمه للغة الإنجليزية. وفي العقد الأخير من القرن الثامن عشر، اعتمد معجم أكسفورد للغة الإنجليزية على جمع كمية هائلة من النصوص كأساس لتأليف المعجم؛ حيث قُدِّمت تلك النصوص إلى أكثر من ثمانمائة متطوع قاموا بقراءتها واستخراج أمثلة عن الكلمات التي سبق تحديدها لهم وكتابة تلك الأمثلة في بطاقات استشهاد ورقية. إلا أن هذه الطريقة في العمل تختلف كثيرًا عن المنهجية المُتَّبَعة حاليًا. أولاً: لأن تلك النصوص المبكِّرة لتكوين الذخائر اللغوية لم تكن مُمَثِّلة للواقع اللغوي، بالإضافة إلى أن هؤلاء المتطوعين كانوا لا يهتمون إلا بالاستخدامات الخاصة للوحدات اللغوية التي يبحثون عنها، ولم يهتموا بالاستخدامات الشائعة لتلك الكلمات. ثانيًا: لأن

تطور تقنيات الحاسب الآلي في عصرنا الحالي قد أتاح للدراسات المعجمية القائمة على ذخائر لغوية تميزًا غير مسبوق يخرج عن نطاق توقعات السابقين في هذا المجال؛ على سبيل المثال من الممكن تصميم ذخيرة لغوية ذات تمثيل لغوي لمنظومة اللغة بأكملها، كأن تحتوي مواد شفهية بكمية كافية، والقدرة على جمع كميات هائلة من النصوص اللغوية وتخزينها وإدارتها، بالإضافة إلى إتاحة أدوات بحث وإحصاء سهلة الاستخدام يمكن الاعتماد عيلها، وغير ذلك من التطبيقات. ومقارنةً بالأسلوب اليدوي، يستطيع الحاسب الآلي البحث عن جميع الأمثلة التي تحتوي كلمة معينة داخل ذخيرة لغوية يبلغ حجمها عشرات الملايين من الكلمات بسهولة شديدة ودون أن يغفل مثالاً واحدًا، بالإضافة إلى قدرته على توليد ملف مُفَهْرَس لتلك الأمثلة. بالإضافة إلى ذلك يستطيع الحاسب الآلي أن يستخدم أساليب أكثر تعقيدًا من تلك التي يستخدمها الإنسان في تحليل أية علاقة من علاقات الارتباط بين الكلمات. على سبيل المثال عندما يصل عدد مرات ظهور كلمة معينة داخل الذخيرة اللغوية إلى آلاف المرات، إذا تم الاعتماد على الإنسان في عَدْ هذه المرات، بالإضافة إلى عمل إحصاء للكلمات التي تظهر مع تلك الكلمة يمينًا ويسارًا في حدود أربع كلمات وتصنيفها، فإن هذا العمل سيكون في إطار المستحيل. أما الآن فيمكن إنجاز هذه المهمة في دقائق معدودة باستخدام أحد أجهزة الحاسوب.

ولذلك، فإن استخدام الحاسب الآلي بدعم من ذخيرة لغوية ذات نطاق واسع، من الممكن أن يساهم في إجراء بحوث استقصائية أكثر عمقًا وأكثر تعقيدًا حول مجموعة الإشكاليات التي يهتم ببحثها علم المعجمية، وهذا ما لم يكن متاحًا من خلال العمل اليدوي فيما مضى.

مبكرًا في الستينيات من القرن العشرين، صممت دار نشر هيريتاج (Heritage) الأميركية ذخيرة AHI[73] لتأليف معجم لطلاب المرحلة الإعدادية. ولكي تضم هذه الذخيرة الكلمات التي يستخدمها الطلاب بصورة حقيقية؛ أي لكي تكون هذه الذخيرة مُمَثِّلَة للواقع اللغوي الذي صُمِمَت من أجله، تم تصميم هذه الذخيرة بعناية فائقة فيما يتعلق بطريقة جمع عينات النصوص اللغوية. فقد قام مؤلفو الذخيرة بإجراء دراسة استطلاعية دقيقة وواسعة النطاق لتحديد المواد التي يتحتم على الطلاب الأميركيين قراءتها وتلك التي يُقبِلُون على قراءتها في إطار ١٠٤٥ نوع من المطبوعات الرسمية، ثم تم اختيار عينات عشوائية من تلك المواد في حدود عشرة آلاف عينة، على ألا تقل العينة الواحدة عن ٥٠٠ كلمة، ليصل بذلك حجم الذخيرة إلى خمسة ملايين كلمة. بالإضافة إلى ذلك، فإن تلك النصوص تنتمي إلى ٢٢ موضوعًا؛ ومن ثَمَّ أمكن التوصل إلى عمومية استخدام الذخيرة. وعلى هذا الأساس، لم تكتفِ دار نشر هيرتاج بإصدار معجم طلاب المرحلة الإعدادية فقط، بل نشرت في عام ١٩٧١ نتيجة الدراسة الإحصائية عن معدلات استخدام الكلمات داخل ذخيرة AHI؛ التي قام بها العالم كارول وفريقه البحثي[74].

أما معجم كولينز كوبويلد (ins CobuildColl) الكبير للغة الإنجليزية[75] الذي اشتُهِرَ أنه أول معجم على مستوى العالم يتم تأليفه باستخدام الحاسب الآلي، فقد تم الانتهاء من إصداره عام ١٩٨٧ بدعم من ذخيرة قاعدة البيانات اللغوية الدولية الصادرة بالتعاون بين دار نشر كولينز

وجامعة برمنجهام المسماة اختصارًا بذخيرة كوبويلد COBUILD(١)؛ التي وصل حجمها إلى عشرين مليون كلمة. وقد تم بناء الذخيرة بتمويل من دار نشر كولينز، أما إدارة تحرير المعجم وتصميم الذخيرة فقد تولى مسئوليته العالم سنكلير من جامعة برمنجهام البريطانية. وقد تم الاهتمام بشدة في أثناء تصميم ذخيرة كوبويلد أن تكون جميع المجالات مُمَثَّلة داخل الذخيرة، وأن تكون الذخيرة مناسبة لمتطلبات العمل المعجمي، وقد ظهر ذلك من خلال النقاط التالية:

(١) تشغل اللغة التحريرية نسبة ٧٥% من حجم الذخيرة، واللغة الشفهية نسبة ٢٥%.

(٢) تُمَثِّل المادة اللغوية للذخيرة اللغة الإنجليزية القياسية، ولا تشمل اللغة الدارجة. وتُمَثِّل اللغة الإنجليزية البريطانية بنسبة ٧٠% واللغة الإنجليزية الأميركية بنسبة ٢٥% أما باقي المناطق الأخرى فتُمَثِّل بنسبة ٥%.

(٣) تُمَثِّل المادة اللغوية استخدامات اللغة الإنجليزية المعاصرة فيما بعد عام ١٩٦٠، ويُراعى أن تكون المادة اللغوية حديثة قدر الإمكان.

(٤) لا تشمل الذخيرة المواد الشعرية والمسرحية والمواد العلمية.

(٥) الاقتصار على جمع أعمال الكتاب فوق الـ ١٦ عامًا، على ألا تقل نسبة الكاتبات عن ٢٥%.

---

(1) (Collins Birminghan University International Language Database)

(٦) المادة التي يتم جمعها ليست عينات أو أجزاء من النصوص، بل نصوص كاملة أو مختارات من روايات كاملة؛ وذلك لمصلحة إجراء تحليل لغوي عن العلاقات البنائية بين أجزاء النص.

وقد كان إصدار هذا المعجم بمثابة نُصُب تذكاري في عالم المعجمية، وكانت أهم ميزة له اعتماده على الواقع اللغوي في اختيار الكلمات واستخداماتها وتقديم الشروح لها وغير ذلك من المحاور الأساسية في العمل المعجمي. وقد صَرَّحَ العالم سنكلير رئيس تحرير المعجم أن هذا المعجم لم يعتمد على أسلوب تأليف الأمثلة التي يتبعه المعجميون، وأن جميع الأمثلة التي استخدمها مُسْتَمَدَّة من مادة لغوية حقيقية. وأن هذه الصفة تُؤَسِّس توجهًا جديدًا سيؤثر تأثيرًا كبيرًا في عالم التأليف المعجمي على المدى البعيد.

وفي عام **1997** أَلَّفَ العلماء التايوانيون خوانغ جو رن، وتشن كه جيان، ولاي تشين شيونغ (معجم الكلمات الكمية للغة الصينية(١) التايوانية في مجال الصحافة)[76]، ويُعتبر هذا المعجم أول مُعجم للغة الصينية يعتمد في تأليفه على ذخيرة لغوية. وينقسم هذا المعجم إلى جزأين رئيسيين: الجزء

(١) الكلمات الكمية في اللغة الصينية عبارة عن كلمة تفصل بين العدد والمعدود، وغالبًا ما تتغير الكلمة الكمية بتغير المعدود مثل "三个学生" بمعنى "ثلاثة طلاب"، و"四本书" بمعنى "أربعة كتب"، وهي ظاهرة لغوية تتميز بها اللغة الصينية عن غيرها من اللغات، وغالبًا لا تظهر هذه الكلمات في النص الهدف المترجم عن اللغة الصينية، واللغة العربية بها مثل هذا النوع من الكلمات ولكن على نطاق ضيق مع بعض الكلمات التي تستحضر كلمة كمية ثابتة كأن نقول: "ثلاثة رءوس من الماشية"، أو عندما تُعَبِّر الكلمة الكمية عن وعاء يحتوي الكلمة موضع العد كأن نقول: "خمسة أكواب من العصير". (المترجم)

المتعلق بالكلمات الكمية والجزء المتعلق بتراكيب الكلمات الكميـــة. يتنـــاول الجزء المتعلق بالكلمات الكمية بالشرح استخدامات الكلمات الكمية، بالإضافة إلى أنواع الأسماء التي يمكن أن تأتلف معها في تركيب. وينطلق الجزء الذي يتناول تراكيب الأسماء مع الكلمات الكمية من الأسماء في اللغـــة الـــصينية، مُعَبِّرًا عن جميع الكلمات الكمية التي يمكن أن تُستَخدم مع تلك الأسماء. وفيما يلي نعرض مثالاً من هذا المعجم(١):

法 [١] 方法。 方法，办法，作法，手法，用法，写法，疗法，玩法，演算法…。指方法或方式。 [一般] 个，项，套。[种类] 样，式。 看法，说法，想法，讲法……。指意见。[一般] 个，项，点。 [种类] 派，样，式。 【辨析】我们可以说 "这一点看法，这一点说法，这一点想法",但是不能说 "这一点讲法"。 [2] 规律。 宪法，劳动法，刑法，民法，交易法，选举法，国安法，著作权法，保育法，国际法，军法，税法……。指各种法律。通常不配量词。 【辨析】"宪法"还可以说"一部宪法"。法律条文的内

(١) سيتم عرض المثال باللغة الصينية كما جاء بالكتاب المصدر، يعقبه ترجمة لمـــا ورد به من معلومات حتى يحتفظ المثال بشكله الأصلي. (المترجم)

容是依
"条，项，款"编列，如"民法第一百八十条第一项第二款，公司法第四百一十九条第一项第五款"。

语法，文法，句法……。指语文的规律。

[一般]套，条，个。

[3] 能力。

枪法，剑法，箭法，刀法，指法……。[一般] 套，条，个。

[种类] 派，式。

佛法，魔法。通常不搭配量词。

书法。[一般]幅，张，篇。

【辨析】"书法"除了和上述量词搭配之外，还有"他的这一手法写得真好"这样的说法。

法

[1] بمعنى طريقة أو منهجية، مثل:

方法，办法，作法，手法，用法，写法，疗法，玩法，...演算法

بمعنى: طريقة، أسلوب، طريقة عمل، مهارة يدوية، طريقة استخدام، أسلوب كتابة، طريقة علاج، طريقة لعب، طريقة حساب، وغير ذلك من الكلمات التي تعبر عن الطريقة أو الأسلوب.

[في العموم] تُستَخدَم الكلمات الكمية: 套，项，个، وتُستَخدَم الكلمات الكمية: 式，样，派 عند التعبير عن فئة أو نوع.

أو مثل:

看法，说法，想法，讲法……。指意见。

بمعنى: وجهة نظر، طريقة حديث، طريقة تفكير، حُجَّة، وغير ذلك من الكلمات التي تُعبِّر عن الرأي.

[في العموم] تُستَخدَم الكلمات الكمية: 点，项，个، وعند التعبير عن فئة أو نوع تُستَخدَم الكلمات الكمية: 式，样.

【تعليق】 يمكننا استخدام الكلمة الكمية " 点" مع الكلمات "看法，说法，想法" مثل:

"这一点想法，这一点说法，这一点看法"

بمعنى "وجهة النظر هذه، هذه الإفادة، هذه الطريقة في التفكير"

ولكن من الخطأ أن تُستَخدَم مع كلمة "讲法" بمعنى "حُجَّة" فلا يمكن أن نقول "这一点讲法".

[٢] بمعنى قانون، مثل:

宪法，劳动法，刑法，民法，交易法，选举法，国安法，著作权法，保育法，国际法，军法，税法……

بمعنى: القانون الدستوري، قانون العمل، القانون الجنائي، القانون المدني، القانون التجاري، قانون الانتخابات، قانون الأمن العام، قانون حقوق المؤلف، قانون رعاية الطفل، القانون الدولي، القانون العسكري،

قانون الضرائب... وغير ذلك من الكلمات التـــي تعبـــر عـــن أنـــواع القوانين. وهذه الكلمات غالبًا ما تَستَغنِي عن الكلمات الكمية.

【تعليق】كلمة "宪法" بمعنى "القانون الدســـتوري" لا يُـــستَخدم معها كلمة كمية فلا يمكن أن نقول "一部宪法"، أما نصوص القـــانون فيمكن ترتيبها من الأكبر فالأصغر فتُـــستَخدَم معهـــا الكلمـــات الكميـــة "条，项，款" بمعنى "مادة، بند، نقطة" فمثلا نقول:

"

民法第一百八十条第一项第二款，公司法第四百一十九条第一项第五款"

بمعنى "النقطة الثانية من البند الأول من المادة المائة والثمانين من القانون المدني، والمادة الخامسة من البند الخامس من المادة الأربعمائة وتسعة عشر من قانون الشركات"

أو مثل:

语法，文法，句法……

بمعنى قواعد اللغة، قواعد الأدب، القواعد النحوية، وغير ذلك من الكلمات التي تُعبّر عن قواعد اللغة والأدب.

[في العموم] تُستَخدَم الكلمات الكمية: 套，条，个

[٣] بمعنى مهارة أو موهبة أو قدرة، مثل:

枪法，剑法，箭法，刀法，指法……

بمعنى: الرماية، المبارزة بالسيف، مهارة تصويب السهام، مهارة استخدام السلاح الأبيض، مهارة العزف على الآلات الوترية،...

[في العموم] تُستَخدَم الكلمات الكمية: 个, 条, 套، وعند التعبير عن فئة أو نوع تُستَخدَم الكلمات الكمية: 式, 派.

أو مثل:

佛法，魔法……

بمعنى المذهب البوذي، أساليب السحر، وهذه الكلمات لا تستخدم معها كلمات كمية.

أما كلمة "书法" بمعنى "فن الخط" فتُستَخدَم معها كلمات كمية.

[في العموم] تُستَخدَم الكلمات الكمية 篇, 张, 幅.

【تعليق】 كلمة "书法" بمعنى "فن الخط" يمكن أن تُستَخدَم معها الكلمة الكمية "手" بالإضافة إلى الكلمات الكمية سالفة الذكر، فنقول: "他的一手书法写得真好" بمعنى "إنه يكتب هذا النوع من الخط بشكل رائع".

وكانت الخطوات التي تم اتباعها لتأليف هذا المعجم التايواني المتخصص في تراكيب الأسماء مع الكلمات الكمية كما يلي:

(١) جمع الأمثلة التي تمثل التصاحبات اللغوية بين الأسماء والكلمات الكمية من الذخيرة.

(٢) تصنيف البيانات التي تم الحصول عليها من الخطوة السابقة حسب معدلات التكرار.

(٣) اختصار البيانات السابقة وفقًا لتصنيفات علماء اللغة ومؤلفي المعاجم.

ولذلك فإن من أهم مميزات هذا المعجم ما يلي:

(١) عدم الاعتماد على استخراج المعلومات من المعاجم السابقة.

(٢) عدم الاعتمام فقط على الرؤية الذاتية للقائمين على تحرير المعجم في كتابة مواد المعجم.

(٣) استخراج الأمثلة اللغوية موضع الاستشهاد بعددها الكبير بشكل مباشر من ذخيرة لغوية، ثم تحليلها وتصنيفها من خلال وجهة نظر علماء اللغة.

وهذا المعجم لم يُقَدِّم فقط وصفًا حقيقيًّا لظروف الاستخدام اللغوي لهذه الظاهرة، بل تَعَمَّق بصورة أكبر وقَدَّمَ شرحًا لأساليب استخدام كل مادة من مواد المعجم، على سبيل المثال، قَدَّمَ المعجم تحت كل مادة عرضًا يفيد بأكثر الكلمات والجمل شيوعًا وتمثيلا في الواقع اللغوي. وتكمن مناطق الاختلاف بين هذا المعجم والمعاجم السابقة في أن الكلمات والأمثلة المُستَخْدَمة ليست من تأليف محرري المعجم، بل تم اختيارها من ذخيرة لغوية كبيرة الحجم مشفوعة بمعدلات التكرار، ثم معالجتها من وجهة نظر هيئة تحرير المعجم وفقًا للقواعد التي أشرنا إليها. ولذلك فإن هذا المعجم لا يمكن أن يختلف عليه اثنان من حيث قوة التأثير وثراء المادة اللغوية.

مُجْمَل القول: إن استخدام منهجية الذخائر اللغوية من شأنه أن يساهم في تطوير الدراسات في علم المعجمية على النحو التالي:

(١) كم عدد المعاني المختلفة للكلمة الواحدة؟

حيث تُعتَبَر هذه القضية من محاور اهتمام علم المعجمية؛ لأن علم الذخائر اللغوية من شأنه أن يُقَدِّم استقراءً للمعاني المتشابهة أو المختلفة للكلمات في السياقات المختلفة من خلال مادة لغوية حقيقية واسعة النطاق، ومن ثَمَّ لا تكون هناك حاجة إلى الاعتماد على الأمثلة التي يؤلفها المعجميون حسب خبراتهم ووجهات نظرهم الخاصة.

(٢) معدل تكرار ظهور كلمة معينة

هذا النوع من الدراسات الاستطلاعية داخل الذخائر اللغوية هو ما يُطلَق عليه إحصاء معدل تكرار الكلمات، وهذا الإحصاء من شأنه أن يُظْهِر لنا درجة شيوع الكلمة، الأمر الذي يجعلنا نميز بين الكلمات شائعة الاستخدام والكلمات النادرة. وهذه المعلومات لها قيمة مرجعية مهمة في تحديد ما إذا كانت الكلمة ستَرِد في المعجم أم لا، أو في تأليف مقررات تعليم اللغات أو في صناعة المعاجم الإلكترونية التي يُعتَمَد عليها في تطوير أنظمة المعالجة الآلية للغات الطبيعية وغير ذلك من المجالات.

(٣) ما الكلمات التي دائمًا ما تتصاحب مع كلمة ما؟

وهذا ما يُطلَق عليه دراسة التصاحبات اللغوية بين الكلمات. فالعالم اللغوي فيرث (**Firth**) كانت له مقولة شهيرة هي: "معانى الكلمات لا تُفْهَم إلا من خلال التصاحبات"[77]. والمقصود بذلك هو أن المعاني المختلفة لكلمة ما لا يمكن تمييزها إلا من خلال التصاحبات التي تظهر مع تلك الكلمة.

وانطلاقًا من وجهة النظر هذه لا يتم تمييز المعاني المختلفة لكلمة ما فقط (الإشكالية الأولى فيما سبق)، بل يمكن تَعَلُّم الاستخدامات المختلفة لتلك الكلمة، وكل ذلك لا يتم التوصل إليه إلا من خلال أنماط الاستخدام والعلاقات التصاحبية بين الكلمات المختلفة. وتكشف لنا النظرة العامة على الإنتاج المعجمي للغة الإنجليزية المنشور بعد تسعينيات القرن العشرين أن جميع تلك المعاجم بلا استثناء قد تم على أساس استقرائي لذخيرة لغوية واسعة النطاق. ولا يَقْتَصِر الأمر عند هذا الحد، بل إن أحدث مُعْجَمَيْن صدرا عن دار نشر لونجمان (manLong) هما: (معجم لونجمان للغة الإنجليزية المعاصرة) (الطبعة الثانية)[78]، و(معجم لونجمان لتداعيات المعنى باللغة الإنجليزية) (Longman Language Activator) (١٩٩٣)[79]، وهذان المعجمان يهتمان بصورة كبيرة بالدور الواضح الذي تلعبه التصاحبات اللغوية في شرح اللغة وتوليدها. وبالنسبة إلى المستخدمين الذين يتعلمون اللغة الإنجليزية باعتبارها لغة أجنبية فإن هذه الطريقة في تأليف المعاجم لها أهمية كبرى؛ ولذلك يلجأ محررو المعاجم إلى جمع كميات كبيرة من الأمثلة عن التصاحبات اللغوية، بالإضافة إلى خروجهم عن المألوف في صناعة المعاجم وجمعهم أمثلة عن التعبيرات اللغوية التي تمثل الاستخدامات الثابتة للكلمات، ويُطْلِقون عليها اسم الكلمات التعبيرية (phrase word) وفي مجال الدراسات المتعلقة بعلم اللغة الحاسوبي والمعالجة الآلية للغات الطبيعية تُعْتَبَر عملية إزالة اللبس الدلالي بين الكلمات؛ التي يُطْلَق عليها اختصارًا Word)WSD Sense Disambiguation) من الموضوعات المُتَّفَق عليها أنها ذات صعوبة خاصة. بالإضافة إلى أن استقراء التصاحبات اللغوية على نطاق واسع يُعْتَبَر بمثابة مقدمة ابتكارية لحل هذه المشكلة.

(٤) كيف يؤثر كل من المجال اللغوي والفترة التاريخية واللغة الدارجة وغير ذلك من العناصر غير اللغوية على نماذج استخدام الكلمات؟

هذا الاستقراء من شأنه أن يساهم في فروق استخدام الكلمات بين المجالات المختلفة، أو عملية تطور الكلمات مع تغير الزمن.

وفيما يلي نقوم بتقديم النتيجة التي توصل إليها العالم بيبر (D. Biber) في أثناء استقراء كلمة DEAL (مكتوبة بالحروف الكبيرة للتعبير عن أنها إجمال لجميع الحالات الصرفية التي تنشأ عن تلك الكلمة)[80].

أولاً: استقراء معدل تكرار كلمة DEAL

للاطلاع على جدول معدلات التكرار لكلمة DEAL في ذخيرة لوب؛ التي يبلغ حجمها عدد مليون كلمة انظر الجدول ٤-١.

جدول (٤-١): جدول معدلات تكرار كلمة DEAL وغيرها من الكلمات في ذخيرة لوب

DEAL	290	182	Deal
OF	35749	52	Dealing
THE	2817	25	Deals
HE	9068	31	Dealt
I	7778	290	الإجمالي
MAKE	2417		
SIGH	16		
APPROACH	185		
LOOK	500		

والجدول رقم ٤-٢ يَعْرِض توزيع معـدلات تكـرار كلمـة DEAL باعتبارها اسمًا سواء كانت في صيغة المفـرد أم الجمـع موزعـة حـسب المجالات في ذخيرة لوب.

جدول (٤-٢): توزيع معدلات تكـرار كلمـة DEAL
على المجالات في ذخيرة لوب

المجال	عدد الكلمات التقريبي داخل المجال	عدد مرات التكرار المبدئي كلمة DEAL	عدد مرات التكرار الموحد (واحد لكل عشرة آلاف كلمة)
التقارير الإخبارية	٨٨،٠٠٠	١٤	٩ .١٥
تعليق ونقد	٣٤،٠٠٠	٤	٨ .١١
موضوعات اجتماعية	٥٤،٠٠٠	٤	٤ .٧
ديانات	٣٤،٠٠٠	٥	٧ .١٤
علوم	١٦٠،٠٠٠	١٦	٠ .١٠
عادات شعبية	٨٨٠،٠٠٠	١١	٥ .١٢
مراسلات	١٥٤،٠٠٠	٢٤	٦ .١٥
قصص عامة	٥٨،٠٠٠	٥	٦ .٨

ونظرًا إلى كبر حجم المخرجات في كل تصنيف للمادة اللغوية، فـإن الأرقام الأولية التي يتم حسابها على معدلات تكرار كلمة **DEAL** لا يمكـن الاعتماد عليها بشكل مباشر في مقارنة ارتفاع معدل تكرار هـذه الكلمـة أو انخفاضها في المجالات المختلفة. ولذلك يتم تنسيب هذه المخرجات الرقميـة إلى وحدة المائة ألف (10^5) مع اعتبار حجم المادة في كل مجـال لحـساب

المخرج النهائي. وعلى فرض أن الرقم الأولي لعدد مرات التكرار يساوي m، وحجم ذخيرة التصنيف يساوي M، والرقم النسبي المشير إلى قيمة التكرار يساوي n، فيمكن حساب تلك القيمة من المعادلة التالية: على سبيل المثال لو تم الحساب على مجال التقارير الإخبارية بذخيرة لوب LOB، تكون المعادلة كما يلي:

$$n = \frac{m}{M} \times 10^5$$

وبما أن نطاق ذخيرة التصنيف M = ٨٨٠٠٠ كلمة، فإن الرقم النسبي لكلمة DEAL هو:

$$n = \frac{14}{88000} \times 100000 = 15.9$$

ومن هذه الأرقام الإحصائية يمكننا أن نلاحظ أن عدد مرات ظهور كلمة DEAL في التصنيفات الثمانية لذخيرة لوب بها ما هو أكبر وما هو أقل من ٥ في حدود أربعة تصنيفات؛ حيث كان عدد مرات تكرار كلمة DEAL في تصنيف المراسلات هو الأعلى حيث وصل إلى ٢٤ مرة. وهذا يدل على أن عدد مرات تكرار كلمة DEAL باعتبارها اسمًا في ذخيرة لوب التي يصل حجمها إلى مليون كلمة ما زال صغيرًا إلى حد كبير. وعلى الرغم من ذلك فإن الأرقام النسبية لعدد مرات تكرار هذا الاسم في التصنيفات المختلفة يُظهِر أن هناك اختلافًا في نسب التوزيع، على سبيل المثال نجد أن الرقم النسبي لعدد

مرات تكرار هذه الكلمة داخل التخصصات الثلاثة (المراسلات، والتقارير الإخبارية، والديانات) تُقَدَّر بضعف القيمة في مجال الموضوعات الاجتماعية.

والجدول ٤-٣ يُبَيِّن النتيجة الإحصائية لعدد مرات تكرار كلمة **DEAL** في ذخيرة أكبر في مجالي الروايات والنصوص العلمية كل على حدة، وهذه الذخيرة عبارة عن عينات تم اختيارها من ذخيرة لونجمان- لانكاستر **Longman Lancaster** في حدود أربعة ملايين كلمة، وقد تم عمل الإحصاء على كل من حالتي الاسم والفعل لكلمة **DEAL**.

جدول (٤-٣): يبين الجدول التوزيعي لعدد مرات تكرار كلمة DEAL في مجالين مختلفين وفقًا لنصوص ذخيرة لونجمان- لانكاستر Longman Lancaster

المجال	العدد التقريبي للكلمات داخل المجال	عدد مرات تكرار كلمة DEAL		الرقم النسبي لعدد مرات التكرار منسوبًا إلى المليون	
		اسم	فعل	اسم	فعل
روايات	٢،٠٠٠،٠٠٠	٢١٧	١٢٧	١٠٧	٦٣
نصوص علمية	٢،٠٠٠،٠٠٠	١٤٩	٣٥٥	٧٤	١٧٦
الإجمالي	٤،٠٠٠،٠٠٠	٣٦٦	٤٨٢	٩٠	١١٩

وهذا الجدول الإحصائي يوضح حقيقة في غاية الأهمية؛ وهي أنه على الرغم من أن إجمالي الأرقام النسبية تشير إلى أن صيغة الفعل من كلمة **DEAL** أكثر شيوعًا من صيغة الاسم (١١٩ في مقابل ٩٠)، نجد أن معدل استخدام الاسم أكثر شيوعًا من الفعل في تصنيف الرويات (١٠٧ في مقابل ٦٣)، أما في مجال النصوص العلمية فنجد أن الوضع عكس ذلك؛ حيث كان

استخدام صيغة الفعل من كلمة **DEAL** أكثر من ضعف صيغة الاسم (١٧٦ في مقابل ٧٤).

تشير نتيجة الاستقراء السابقة إلى أن الرقم الإحصائي الإجمالي للمشاهدات داخل الذخيرة بأكملها لا يمكن أن ينطبق دائمًا على المشاهدات المُسْتَخْرَجة من التصنيفات الفرعية للذخيرة. وعلى العكس من ذلك فإن النتيجة التي يتم الحصول عليها من تصنيف فرعي لا تعبر عن باقي المجالات، بل لا يمكن تعميمها لتُعَبِّر عن القانون العام الذي يحكم لغة من اللغات. وبعبارة أخرى، فإن معدلات تكرار كلمة من الكلمات وأنماط استخدامها تُعَبِّر بدرجة كبيرة فقط عن المجال اللغوي الذي تمت عليه عملية الإحصاء. ومن هنا يرى بيبر أن النتيجة الإجمالية التي يتم عملها على لغة ما (اللغة الإنجليزية على سبيل المثال) لا يمكن أن تكون نتيجة استرشادية؛ لأن هذه النتيجة عبارة عن معالجة لمجموع المشاهدات التي تم جمعها عن معدلات الاستخدام في عدد من المجالات. فمن ناحية يرجع هذا إلى أن النتيجة الإجمالية لا تكون سليمة؛ بسبب تحول الأداء اللغوي الذي يحدث للغة من مجال إلى مجال. ومن ناحية أخرى يمكننا القول: إن اللغة التي يُعَبِّر عنها هذا النوع من النتائج الإجمالية غير موجودة على أرض الواقع.

ثانيًا: الاستقراء على مستوى الدلالة

في الغالب يتم عمل الاستقراء على مستوى الدلالة بدءًا من الجدول السياقي للكلمات (**Key Word in Context**) المسمى اختصارًا بـ **KWIC**)؛

حيث يقدم هذا الجدول السياقي جميع الحالات التي ظهرت فيها كلمة معينـــة، مع السياقات المصاحبة لها. ولكن بالنسبة إلى الكلمات متعددة الدلالة ينبغـــي الاعتماد على المجهود البشري في للتعرف على معنى الكلمة موضع البحث (يُطلَق عليها الكلمة الهدف) في كل سجل مُفَهْرَس أو كل مثال من الأمثلـــة، وهذه من أصعب المهام التي يمكن أن يقوم بها العنصر البشري. على سبيل المثال نجد أن كلمة **DEAL** ظهرت ٢٠٠٠ مرة داخل الذخيرة المكونة مـــن مليون كلمة. أما بالنسبة إلى كلمة من الكلمات شائعة الاستخدام فقـــد تـــصل النتائج إلى ما يزيد عن عشرة آلاف سجل. وإذا أردنا أن نستخرج الأنمـــاط الدلالية لتلك الكلمة من بين هذا الكم من النتائج، نحسَب أن ذلك من الأعمال المستحيلة التي لا يستطيع أن يقوم بها عنصر بشري. لذلك فقد اختار العـــالم بيبر طريقًا مختلفًا للاستقراء الدلالي عن طريق التصاحب. إن ما يُطلَق عليه الكلمات المتصاحبة يُقْصَد به تلك الكلمات التي تَرِدْ دائمًا بمـــصاحبة الكلمـــة الهدف داخل النصوص اللغوية. إن هذه الطريقة التي تعتمد علـــى اســـتقراء التوزيع الدلالي للكلمة تقوم على أساس الفرض التالي: كل مجموعـــة مـــن الكلمات المتصاحبة يقتصر تصاحبها مع الكلمة الهدف في إطار معنًى واحد لها فقط. ولذلك فمن خلال تحليل مجموعة الكلمات المتصاحبة التـــي يكثـــر ظهورها مع الكلمة، يمكننا أن نُميِّز بشكل فعال معنى أو معاني تلك الكلمة.

والجدول ٤-٤ يظهر لنا التصاحبات الأكثر ظهورًا مع كلمة **DEAL**. والذخيرة التي تم الاعتماد عليها مُستَخْرَجة مـــن تـــصنيفين داخـــل ذخيـــرة لونجمان لانكاستر **Longman Lancaster**: نصوص علميـــة فـــي حـــدود

مليونين وسبعمائة ألف كلمة، وروايات في حدود ثلاثة ملايين كلمة. ويُقْصَد بالتصاحب الأيسر تلك الكلمات التي ترد قبل الكلمة الهدف مباشرة، مثل كلمة "good" في التصاحب "good deal"؛ أما التصاحب الأيمن فيُقْصَد بــه تلــك الكلمات التي ترد بعد الكلمة الهدف مباشرة، مثل كلمة " of" في التــصاحب "deal of".

وقد أظهر الجدول ٤-٤ أن أكثر الكلمات التي تتصاحب مــع الاســم DEAL من جهة اليسار هي كلمة "great"؛ (ظهرت بمعدل ٤٥ مرة في كل مليون كلمة)، يلي ذلك كلمة "good" (ظهرت بمعدل ٢٣ مرة). وفي الحقيقة أن هذا يمثل مجموع أمثلة يصل إلى ١٨٥ تصاحبًا من بين ١٩٦ حالة ظهور لهذه الكلمة في هذا التصنيف. وتلي ذلك من حالات التصاحب الأيسر كلمــة "package"، وكلمة "that"؛ حيث وردت كل منهما مرتين فقــط أي بمعــدل ٠,٧ كل مليون كلمة.

جدول ٤-٤ الكلمات دائمة التصاحب مع كلمة **DEAL**

المجال		كلمة التصاحب	الإحصاء المبدئي	الإحصاء منسوبًا إلى مليون كلمة
نصوص علمية (٢٧٠٠٠٠٠ كلمة)	تصاحب أيسر	Great	12	45
		Good	63	23
	تصاحب أيمن	Of	106	39
		More	18	7
		In	8	3
		To	8	3
روايات (٣٠٠٠٠٠٠)	تصاحب أيسر	Great	122	40
		ood	84	28
		The	24	8
		Big	10	3
	تصاحب أيمن	Of	84	28
		To	22	7
		About	15	5
		More	10	3
		With	9	3

وتشير تلك النتيجة الاستقرائية إلى أن كلمة **DEAL** في حالتها الاسمية يكثُر ورودها مع التصاحبات "**good/great deal**" في النصوص العلمية، وفي تلك الحالة يكون معنى "**Idea**" هو "كمية"، أو "صفقة". وإذا نظرنا إلى التصاحبات اليمنى سنجد أن كلمة "of" هي أكثر الكلمات المتصاحبة مع الكلمة الهدف من جهة اليمين؛ حيث وردت بمعدل ٣٩ مرة في كل مليون كلمة، وهي نسبة تفوق كثيرًا نسبة الكلمة التي وردت في الترتيب التالي لها (وردت كلمة **more** مع الكلمة الهدف بمعدل سبع مرات). ومن هنا يمكننا أن

نستخلص أن كلمة DEAL عندما تُستخدَم اسمًا فإن نمط التصاحبات المقترن معها يكون على الصيغة "a good/great deal of"، ويتضح من ذلك أيضًا أن المعنى الأكثر تداولاً لكلمة DEAL هو "كمية"، أو "تبادل تجاري". وبالإضافة إلى ذلك فإن الجمع بين نتيجة الاستقراء السابقة وجدول الفهرسة من شأنه أن يُحقِّق لنا مستوى أكثر واقعية للحكم على المعاني الأكثر شيوعًا لكلمة DEAL في حالتها الاسمية. على سبيل المثال، يعبر جدول الفهرسة عن أن أكثر الاستخدامات شيوعًا للنمط "good/great deal" هو " a good great deal of work"، و"a good/great deal attention". ليس هذا فقط، بل إن مجموعة الكلمات المتصاحبة مع كلمة DEAL من جهة اليمين تُعبِّر عن أن معناها في تلك الحالة يكون "كمية". على سبيل المثال، الأمثلة التي وردت فيها كلمة التصاحب "more" من جهة اليمين هي " a great deal more tolerance"، و"a great deal more inhibited". أما كلمتا "in"، و"to" عندما تتصاحبان مع كلمة DEAL من جهة اليمين فيعني ذلك أنها ما زالت تعني " كمية"، مثل: "a great deal in common"، و" differ agreat deal in their understanding"، و"a great deal to be desired"، و" a great deal to offer" وما إلى ذلك. والخلاصة أن غالبية معاني كلمة DEAL في حالتها الاسمية في النصوص العلمية تأتي بمعنى "كمية".

أما إذا قارنا بين النصوص العلمية ونصوص الروايات فسوف نجد أن هناك تشابهًا في حالات التصاحب لكلمة DEAL في صيغتها الاسمية، مع وجود بعض الاختلافات الواضحة. فمن ناحية نجد أن الكلمتين الأكثر

تصاحبًا معها من جهة اليمين ما زالت هي "great"، و"good". ففي الحقيقة أن معدل تكرار نمط التصاحب "good /great+deal" في التصنيفين متماثلا تمامًا؛ حيث سَجَّل ٦٨ مرة لكل مليون كلمة. ومن ناحية أخرى علينا أن ننتبه إلى أن هناك ٩٦ حالة من حالات التصاحب داخل تصنيف الروايات لم تستخدم النمط "good/great+deal" بل ظهرت تصاحبات أخرى جهة اليسار مثل "the" التي ظهرت ثماني مرات كل مليون كلمة، وكلمة "big" التي ظهرت ثلاث مرات كل مليون كلمة، هذا بالإضافة إلى وجود سبع كلمات أخرى ظهرت مرة أو مرتين مع كلمة **DEAL** من اليسار.

وهذا يشير إلى أن معنى "كمية" لكلمة **DEAL** في حالتها الاسمية هو المعنى الأكثر شيوعًا في التصنيفين، على الرغم من ظهور العديد من التصاحبات الجديدة الأكثر شيوعًا؛ التي تُعَبِّر عن معاني (أو استخدامات) الكلمة لم تظهر أمثلة لها في النصوص العلمية. على سبيل المثال كلمة "the" عندما تتصاحب مع كلمة **DEAL** يكون معناها "اتفاقية"، مثل: " **part of the deal is ...**"، و"**Isn't that the deal?**"، أما كلمة "big" فعندما تتصاحب مع DEAL يكون معناها "غير مهم"، مثل: "**no big deal**"، و" **what's the big deal?**".

هذا بالإضافة إلى أن هناك العديد من التصاحبات التي لم يمكن حصرها في الجدول السابق؛ بسبب انخفاض معدل تكرارها تعرضت لمعنى مهمّ من معاني كلمة **DEAL** وهو معنى "صفقة"، مثل: "**property deal**."، و"deal **record**" و"**cash deal**"، و"**Land deal**"، وما إلى ذلك.

وتتشابه المشاهدات التي تم الحصول عليها لمصاحبات كلمة DEAL من جهة اليمين بين النصوص العلمية ونصوص الروايات. إلا أن كلمات مثل "about"، و"with" لم ترد في النصوص العلمية.

وعندما تقترن كلمة "about"، و"of" مع كلمة DEAL نجد أنها ما زالت تحمل معنى "كمية"؛ مثل:

"I also knew a great deal about love".

"We both laughed a great deal about this".

وبالإضافة إلى ذلك، نجد أن كلمة "with" عندما تقترن من جهة اليمين مع DEAL يكون معناها "صفقة"؛ مثل:

"I made a deal with the doctors"

"I'll cut a deal with you"

وفي نصوص الروايات، لوحظ أن هناك حالات من المعنى لم تظهر ولو مرة واحدة في النصوص العلمية؛ بمعنى أن هناك بعض التصاحبات اليمنى مثل كلمة "table" التي ظهرت أربع مرات، وكلمة "box" التي ظهرت مرة واحدة؛ حيث كان معناها "مادة خشبية"، وعلى الرغم من أن هاتين الكلمتين لم تظهرا بعدد مرات كبير، فإنهما تعبران عن استخدام آخر لكلمة DEAL في النصوص الروائية.

وقد أجرى العالم بيير مقارنة على النتائج الاستقرائية التي توصل إليها مع عدد من المعاجم الشائعة فوجد أن بعض هذه المعاجم قد أفردت مَدْخَلا واحدًا (entry) لكلمة DEAL، والبعض الآخر قد أُفرِدَ لها أربعة مداخل على

أكثر تقدير. إن أكثر المعاجم أفردت ٤ مداخل لتلك الكلمة. وقد اختلفت المعاجم فيما بينها في عرض المعاني الشارحة لكلمة DEAL؛ حيث تراوحت تلك المعاني ما بين معنيين أو ثلاثة، وبعضها وصل إلى ثلاثين معنى، وقد ترتب على ذلك صعوبة تمييز القارئ للمعاني المختلفة لكلمة DEAL. وفيما يلي نعرض تجميعًا لعدد سبعة معاني تكررت داخل خمسة معاجم إنجليزية عن كلمة DEAL عندما تأتي على الحالة الاسمية:

(١) كمية كبيرة، قَدْر كبير، كثير.

(٢) اتفاقية.

(٣) (في أوراق اللعب) يوزّع أوراق اللعب على اللاعبين.

(٤) (يتلقى) معاملة.

(٥) توزيع.

(٦) خشب الصنوبر.

(٧) تجارة، معاملة.

وقد أوردَت أغلب المعاجم شرحًا لتلك المعاني السبعة، إلا أن هناك معجمين من بين المعاجم الخمسة لم يرد بهما معنى "توزيع"، وهناك معجم لم يرد به معنى "اتفاقية". هذا بالإضافة إلى وجود فرق كبير بين تلك المعاجم في ترتيب شرح المعنى. على سبيل المثال ظهر معنى "كمية كبيرة، وقدر كبير، وكثير" في المادة الفرعية الثانية في معجم وبستر (Webster)، أما في معجم

راندوم هاوس (Random House) فوردت في المادة الفرعية الحادية والعشرين. وقد كانت الملاحظات التي توصل إليها بيير من خلال المقارنة ما يلي:

(١) على الرغم من أن معنى "كثير" هو المعنى الأكثر شيوعًا داخل التصنيفين الذين تكونت منهما الذخيرة موضع البحث، فإن هذا المعنى لم يتصدر قائمة المعاني في بعض المعاجم المنشورة، بل أتى في آخر الشرح؛ حيث احتل الموضع السادس عشر أو الحادي والعشرين.

(٢) تم التوصل إلى المعنى الأكثر شيوعًا لكلمة DEAL من خلال تحليل التصاحبات اللغوية لهذه الكلمة؛ وذلك عندما نستخدم التصاحب "big deal" بمعنى "غير مهم"، وهذا المعنى لا يوجد معجم حتى الآن تمكن من رصده والتعرض له في الشرح.

(٣) جميع المعاجم الخمسة التي تم استطلاعها أوردت معنى "توزيع أوراق اللعب" باعتباره أحد معاني كلمة DEAL، إلا أن هذا المعنى لم يظهر في الذخيرة التي أجرى العلم بيير بحثه عليها. فمن ناحية تشير نتيجة استقراء الذخيرة إلى أن هذا المعنى من المعاني النادرة لكلمة DEAL. ومن ناحية أخرى فإن المتحدثين الأصليين باللغة الإنجليزية (native speakers) يعتبرون أن هذا المعنى من المعاني الخاصة لكلمة DEAL ولذلك فإن تدوينهم لهذا المعنى في معاجمهم يعتبر عملا صحيحًا. ومن هذا المنطلق يمكننا القول: إن ذلك هو

أحد أوجه القصور في منهجية العمل من خلال الـذخائر اللغويـة. ويرجع ذلك إلى أن أسلوب العينات الذي يتبعه علم الذخائر اللغوية في جمع النصوص يتعرض لقدر من التحيز وتحجيم نطاق المـادة اللغوية، وهذا الوجه من أوجه القصور من الصعب تجنبه. لذلك لا بُدَّ من تدخل علماء اللغة في أثناء عمليات استقراء المفردات؛ حيث تسد معلوماتهم وحسهم اللغوي أوجه القصور في منهجية العمل من خلال ذخائر لغوية.

ثالثًا: منهجية الذخائر اللغوية في دراسة التصاحب اللغوي

يظهر التصاحب اللغوي باعتباره أهم الظواهر الواجب دراستها لخدمة مجالات تعليم اللغات للناطقين بغيرها، والترجمة الآلية، والتوليد الآلي للغات الطبيعية وغير ذلك من المجالات. فلماذا نقول: "يرتدي ملابـس"، و"يـضع قبعة"، في حين أنه لا يمكننا أن نقول: "يضع ملابس"، و"يرتدي قبعة" ولماذا في اللغة الصينية يمكننا استخدام الفعل "看" بمعنى يرى أو يشاهد في الأمثلة التالية "看电影" بمعنى يشاهد فيلمًا، و"看球赛" بمعنى يشاهد مباراة كـرة القدم، و"看小说" بمعنى يقرأ رواية، و"看朋友" بمعنى يزور صديقًا، فـي حين أن هذا الفعل عندما ينقل إلى اللغة الإنجليزية سـيترجم إلـى "go/see visit،read،watch،to" على التوالي، وكل هذه مـن المعلومـات الواجـب استيعابها جيدًا حتى يتمكن الدارس من استخدام لغة من اللغات.

قَدَّمَ المُعْجَمي مورتون بنسون (Benson .M) مُؤلِّف معجم بي بي آي للتصاحب اللغوي في اللغة الإنجليزية ((BBI Cominatory Dictionary of English)) (Johne Ben jamins Publishing Co,1986).[81] تعريفـــا للتصاحب اللغوي كما يلي[82]:

"التصاحب اللغوي هو نوع من تكرار الظهور الإلزامي لمجموعة من الكلمـــات معًـــا (A collocation in an arbitrary and recurrent word combination) ".

وقد حدد بنسون من خلال تعريفه خاصيتين من خـــواص التـــصاحب اللغوي وهما أن ظهور التصاحب يكون إلزاميًّـــا، بالإضـــافة إلـــى تكـــرار الظهور.

والمقصود بلزوم التصاحب هو التفريـــق بـــين التـــصاحب اللـــصقي (bound combination) أو التصاحب الحر (free combination). يعبـــر التصاحب اللصقي عن لزوم عملية التصاحب بـــين الكلمـــات المتـــصاحبة. وبعبارة أخرى، يتمتع هذا التصاحب بصفات خاصة ومحددة، وهي أن كـــل تصاحب يضم على أقل تقدير كلمة واحدة تخضع لقدر كبير من الإلزام فـــي أثناء التصاحب مع الكلمات الأخرى، وعلى هذا الأساس لا تكون حرة فـــي عملية التصاحب؛ مثل التصاحب murder commit بمعنى (ارتكب جريمـــة قتل)؛ حيث إن فعل "قتل" أو "commit" لا يمكن أن يقترن إلا مع عدد محدود من الأسماء مكونا تعبيرة فعل ومفعـــول، مثـــل الكلمـــات "crime" بمعنـــى

"جريمة"، و"suicide" بمعنى "انتحار"، و"doing-wrong" بمعنى "فعل مشين"، وما إلى ذلك. ولذلك يتم اعتبار تعبيرة commit murder تعبيرة لصقية. والتعبيرات اللصقية (التصاحبات) تفرضها العادة اللغوية في الاستخدام، ولا تكفي المعلومات النحوية والدلالية لإماطة اللثام عن السبب في حتمية الحديث بمثل هذه الطريقة؛ ولذلك تكون هذه التصاحبات إلزامية، ولا يمكن للمستخدم أن يتوقع حدوثها. وعلى العكس من ذلك، فإن كل كلمة من الكلمات المكوّنة للتعبيرات الحرة يمكنها أن ترتبط بحرية مع كلمات أخرى خارج نطاق تلك التعبيرة، مكونة بذلك تراكيبًا أخرى على نمط التركيب النحوي الحالي نفسه. على سبيل المثال تعبيرة فعل ومفعول على النمط نفسه مثل "condemn murder" بمعنى "يدين جريمة قتل" تنتمي إلى التعبيرات الحرة وذلك لأن فعل condemn يمكن أن يقترن مع العديد من الأسماء مثل "abduction,abortion,abuse of power, acquittal" وما إلى ذلك باعتبارها مفعولا، أما الاسم murder فمن الممكن أن يقترن مع أكثر من مائة فعل مختلف باعتباره مفعولاً مثل: "abhor,accept,acclaim, advocate"، وما إلى ذلك. لذلك فإن هذه التعبيرات لا تتمتع بخصوصية لغوية، وما على دارسي اللغة الإنجليزية باعتبارها لغة أجنبية ثانية إلا أن يتعرفوا على معنى تلك الكلمات وخصائصها النحوية وما يترتب على ذلك من قواعد نحوية، حتى يمكنهم أن يكوّنوا تلك التعبيرات اللغوية بحرية تمكنهم من التواصل اللغوي السليم. وانطلاقًا من هذا المعنى، فإن التعبيرات الحرة ليست إلزامية ويمكن توقع حدوثها.

وفي مجال دراسات التصاحب في اللغة الصينية فقد نُشِر في الصين عدد من معاجم التصاحب اللغوي. إن دراسة التصاحب اللغوي في اللغة الصينية لا يمكنه أن يتجنب أيضًا قضية التمييز بين حدود التصاحب وغير التصاحب اللغوي. يرى قوه تشيان أن معجم (التصاحبات اللغوية بين الكلمات الحقيقية في اللغة الصينية الحديثة)[83] هو أفضل المعاجم التي تناولت قضية التصاحب اللغوي في اللغة الصينية الحديثة. وذلك على الرغم من أن هذا المعجم قد سجّلَ عددًا كبيرًا من التعبيرات اللغوية الحرة على اعتبار أنها تصاحبات لغوية. وقد عرض رئيس تحرير المعجم في مقدمته فِكْرَه في أثناء عملية التأليف والمتمثل في أن يقدم لكل كلمة حقيقية يرغب أي شخص في التعرف عليها ما يلي:

(١) هل يمكن أن تَشْغَل محل المسند إليه في الجملة؟ وإذا كان من الممكن ذلك، فما الكلمات التي يمكن أن تشغل محل المسند معها؟

(٢) هل يمكن أن تشغل محل المسند في الجمل؟ إذا كان من الممكن ذلك، فما الكلمات التي يمكن أن تشغل محل المسند إليه، أو المفعول، أو المكمل معها؟ وما إلى ذلك. وقد شبَّه فريق تأليف المعجم الكلمة أنها مثل المغناطيس، وأن الكلمات التي تنجذب لتأثير المجال المغناطيسي لتلك الكلمة يمكنها أن تُكَوِّن معها تصاحبًا لغويًّا. وعلى هذا يمكننا أن نتصور كيف أن هذا المعجم سوف يجمع بالتأكيد عددًا لا بأس به من التعبيرات الحرة (لا تمثل تصاحبات لغوية) مثل تصاحبات المسند إليه + المسند مثل "经理能干" بمعنى"المدير الكفؤ"، و"工人能干" بمعنى "العامل الكفؤ"، وتصاحبات المسند

+ المفعول به مثل: "称赞小伙" بمعنى "يمدح الشاب"، و"称赞学生" بمعنى "يمدح الطالب"، وما إلى ذلك.

ومن هذا المنطلق، فإن التوصل إلى طريقة إحصائية مناسبة للحكم على ما إذا كان التركيب يمثل تصاحبًا لغويًا أم غير ذلك يُعْتَبَر من القضايا المهمة التي تشغل العاملين في حقل الذخائر اللغوية. كان العالم ياكوف شويكا (Y .Choueka) أول من استخدم الأدوات الحاسوبية في عمل تحليل كمي للتصاحبات اللغوية في الغرب[84] حيث عرَّف هو وفريقُه البحثي التصاحب اللغوي على أنه سلسلة من الكلمات المتجاورة التي تتكرر معًا، وقد نجح في استخراج عدة آلاف من التصاحبات اللغوية الشائعة في اللغة الإنجليزية آليًّا من ذخيرة تصل إلى إحدى عشر مليونًا من الكلمات من جريدة نيويورك تايمز باستخدام الحاسب الآلي؛ مثل: fried chicken Johnson, home run, Magic، وغيرها. وقد كان من عيوب تلك الدراسة عدم التفكير في إمكانية أن تقوم إحدى الكلمات بالفصل بين الكلمتين المتصاحبتين؛ مثل: (decision... make) بالإضافة إلى الخاصية الإلزامية للتصاحب اللغوي. أما العالم كَيَنيت والت تشرتش (Church .K) وغيره من العلماء فقد عَرَّفوا التصاحب اللغوي على أنه عبارة عن كلمتين بينهما علاقة متبادلة، واستخدموا مفهوم "كمية المعلومات المتبادلة" (mutual information) الخاص بنظرية المعلومات في الحكم على القوة الاقترانية بين أي كلمتين في اللغة[85]؛ حيث أجرَوا اختبارًا على ذخيرة من النصوص الإخبارية مكونة من أربعة وأربعين مليون كلمة أُطْلِقَ عليها ذخيرة (Usap Corp). إن القيمة الإحصائية التي تقدمها معادلة

المعلومة المتبادلة يمكنها أن تقدم إلى حد بعيد تفسيرًا للخاصيتين اللتين تم طرحهما آنفًا؛ وهما خاصية الإلزام وخاصية التكرار، هذا بالإضافة إلى أنها لا تتقيد بضرورة أن تكون الكلمتان متجاورتين. ومن عيوب تلك الطريقة عدم الاهتمام بأن يتضمن التصاحب دائمًا خصائص تركيبية نحوية، الأمر الذي تسبب في أن يستخرجوا العديد من أزواج الكلمات من الذخيرة؛ مثل: **doctor-nurse, doctor-bill, doctor-hospital** على الرغم من وجود علاقة دلالية بين تلك الكلمات؛ ولكن نظرًا إلى عدم وجود علاقة نحوية تحكم تلك الكلمات، فلا يمكن النظر إليها باعتبارها تصاحبات لغوية إذا نظرنا إليها بقدر أكبر من الدقة في الحكم. أما برمجية إكستراكت **Xtract** التي صممها العالم فرانك سمادجا (**Smadja .F**) فتُعْتَبَر من أحدث البرمجيات التي تقوم بعمل تحليل كمي للتصاحبات اللغوية بالإضافة إلى أنها-الأكثر تكاملا[86]. ولم يكتف سمادجا بتقديم معادلة تفسر قوة الاقتران بين أزواج الكلمات فحسب، بل استعان بمعادلة لحساب موقع الكلمة ودرجة التشتت في توزيعها. وقد نجحت برمجية إكستراكت **Xtract** في استخراج أمثلة عن التصاحب اللغوي من ذخيرة تصل إلى عشرة ملايين كلمة عن أخبار بورصة الأوراق المالية، وقد وصلت نسبة الدقة في النتائج إلى ما يقرب من ٨٠%.

استخدم العالم سوين ماو سونغ وغيره من جامعة شين خوا الصينية ذخيرة الأخبار التي أنشأتها وكالة أنباء الصين الجديدة شين خوا في الفترة ما بين عامي ١٩٩٠-١٩٩١؛ التي أُطْلِقَ عليها اسم **XH-CORPUS**، في عمل تجربة عن التحليل الكمي للتصاحبات اللغوية في اللغة الصينية، وكان الهدف

من ذلك تقديم مرجعية كَمِّيَة عن التصاحبات اللغوية لعلماء اللغة حتى يمكن تقديم كشاف نصف آلي عن التصاحب اللغوي يصلح لأنماط التواصل اللغوي بين الإنسان والآلة[87]، وفي الصفحات التالية سوف نقدم نتيجة الاستقراء الكمي الذي قام به سوين ماو سونغ عن التصاحب اللغوي.

قوة التصاحب

قام كينيت والت تشرتش وغيره من العلماء باستخدام معادلة كمية المعلومات المتبادلة mi لقياس درجة العلاقة بين أي كلمتين w، و w_i.

$$mi(w,w_i) = \log_2 \frac{p(w,w_i)}{p(w)p(w_i)}$$

المعادلة (٤-١)

حيث تشير $p(w,w_i)$ إلى احتمال ورود الكلمتين w، و w_i في نطاق السياق، أما $p(w)$، و $p(w_i)$ فتشير كلٌ منهما إلى احتمال ورود $p(w)$، و $p(w_i)$ بمفردهما على التوالي داخل الذخيرة.

على فرض أن $p(w)$، و $p(w_i)$ زوج من الكلمات المتصاحبة المُنْتَخَبَة، تعكس القاعدة (٤-١) السابقة مدى إلزامية التصاحب، وخاصية تكرار الظهور.

(١) عندما تكون قيمة $p(w)$، و $p(w_i)$ ثابتة لا تتغير، وكانت درجة تَقَيُّد الكلمتين w، و w_i ببعضهما كبيرة، وكانت فرصة ظهور هاتين الكلمتين مع كلمات أخرى قليلة، بالإضافة إلى أن قيمة

، $p(w)$ أو $p(w_i)$ قليلة، فمن ثَمَّ تكون كمية المعلومـات المتبادلـة بينهما $mi(w,w_i)$ كبيرة. وهذا يعبِّر عن أن كـلا مـن w، و w_i بينهما علاقة اقتران إلزامية قوية. والعكس من ذلك صحيح.

(٢) عندما تزيد عدد مشاهدات w، مع w_i، ويكون احتمال ورود w مع w_i والذي يُرْمَز له بالرمز $p(w,w_i)$ كبيرًا، تزداد تبعًـا لـذلك كمية المعلومات المتبادلة بين w، و w_i؛ التـي يُرمـز لهـا بـ $mi(w,w_i)$. وهذا يُعَبِّر عن زيادة قوة التصاحب بين كل مـن w، و w_i. والعكس من ذلك صحيح.

ويراعى في أثناء اكتشاف أمثلة التصاحب ضرورة أن يأتي التصاحب بين الكلمتين w، و w_i في إطار جملة واحدة. ففي إطار الجملة الواحدة يمكن السماح بأن تتفصل الكلمتان w، و w_i عن طريق مجموعة من الكلمات. على سبيل المثال، تعبيرات مثل "穿衣服" بمعنى "يرتدي ملابس"، و"穿新衣服" بمعنى "يرتدي ملابس جديدة"، و"穿了一件红衣" بمعنى "يرتـدي ملابـس حمراء اللون"، وما إلى ذلك من تعبيرات لغوية أو جمل، نجد أن كلمـة "穿" بمعنى "يرتدي" تتصاحب مع كلمة "衣服" بمعنى "ملابس" وهما كلمتان ينبغي ظهورهما معًا. وبالطبع فإنه كلما بعدت المسافة بين الكلمتين أمكننـا القـول بانحسار قوة العلاقة التصاحبية فيما بينهما. وقد حددت برمجية إكـستراكت Xtract السياق المحيط بالكلمة الإنجليزية عند حدود خمس كلمات قبل تلـك الكلمة وبعدها. وبعبارة أخرى، يمكن القول: إن نافذة الرؤيـة للكلمـة تـم

تحديدها بقيمة ±٥ كلمات. وقد اتبع سوين ماو سونغ طول النافذة نفسه في استقراء التصاحب اللغوي في اللغة الصينية؛ حيث جعل الصيغة $p_j(w,w_i)$ تُعبِّر عن احتمال ظهور الكلمتين w، و w_i معًا عندما يفصلهما عدد من الكلمات يقدر بالمسافة j. على أن تكون $j=-5,-4,-3,-2,-1,1,2,3,4,5$. فعندما تظهر الكلمة w_i على يسار الكلمة w، فإن j تأخذ قيمة سالبة، أما عندما تظهر الكلمة w_i على يمين الكلمة w فإن j تأخذ قيمة موجبة.

فإذا عبرنا عن قوة التصاحب بالصيغة $s(w,w_i)$ فيُمكِننا حساب ذلك من خلال المعادلة التالية التي تقوم على أساس من القاعدة (٤-١) كما يلي:

$$s(w,w_i)=\log_2\frac{\sum_{j=-5}^{+5}p_j(w,w_i)}{p(w)p(w_i)}$$

المعادلة (٤-٢)

فلو فرضنا أن مجموع الكلمات المُكوِّنة للذخيرة هو N، وكانت الصيغة $r_j(w,w_i)$ تُعبِّر عن ظهور الكلمتين w، و w_i وتفصلهما المسافة j، وكانت كل من $r(w)$، و $r(w_i)$ تُعبِّران عن عدد مشاهدات w، و w_i بمفردهما داخل الذخيرة، واستخدمنا التقديرات الأعلى فيمكننا الحساب بصورة منفردة كما يلي:

$$p_j(w,w_i)=r_j(w,w_i)/N$$
$$p(w)=r(w)/N$$
$$p(w_i)=r(w_i)/N$$

ويمكننا إدراج ماسبق مع القاعدة (٤-٢) فنحصل على الصيغة التالية:

$$s(w,w_i)=\log_2\frac{N\sum_{j=-5}^{+5}r_j(w,w_i)}{r(w)r(w_i)}$$

المعادلة (٤-٣)

يصل حجم ذخيرة **XH-CORPUS** إلى عشرة ملايين رمز، وبعد أن تمت عليها عملية تقسيم للكلمات وصل عدد كلماتها إلى $N=7.1\times10^6$ وقد أجرى سوين ماو سونغ دراسة استقرائية على التصاحبين اللغويين "能力，弱" بمعنى "قدرة، ضعيفة"، و"能力，大" بمعنى "قدرة، كبيرة". وفيما يلي بيانات تلك الدراسة الإحصائية:

المجموعة الأولى: "能力，弱"

r_{-3} = (能力，弱) =1，r_1 (能，弱) =3

r_2 (能力，弱) =5，r_j (能力，弱) =0 (j=-5,-4,-2,-1,3,4,5)

r=(能力)=2441, r=(弱)=177

ووفقًا للمعادلة رقم (٤-٣) يتم التعويض في المعادلة كما يلي:

$$s(能力，弱) = \log_2 \frac{7.1\times10^6(1+3+5+7\times0)}{2241\times177} = 7.33$$

المجموعة الثانية:"能力，大"

r_{-5}（能力，大）=6, r_{-4}（能力，大）=4,

r_{-3}（能力，大）=8, r_{-2}（能力，大）=4, r_{-1}（能力，大）=2,

r_1（能力，大）=9, r_2（能力，大）=6, r_3（能力，大）=4,

r_4（能力，大）=6, r_5（能力，大）=5, r（能力）=2241,

r（大）=19913

ووفقًا للمعادلة رقم (٤-٣) يتم التعويض في المعادلة كما يلي:

s（能力，大）=

$$\log_2 \frac{7.1\times10^6(6+4+8+4+2+9+6+4+6+5)}{2241\times19913} = 3.10$$

من خلال درجة قوة التصاحب في المجموعتين نلاحظ أن S（能力，弱）أكبر بكثير من S（能力，大）، ومن هنا يتضح لنا أن المجموعة (能力，弱) تتجه لتكوين تصاحب لغوي بصورة أكبر من (能力，大). وعلى الرغم من أن الكلمتين "能力" و"弱" وردتا معًا في الذخيرة عدد ٩ مرات، وأن الكلمتين "能力" و"大" وردتا معًا عدد ٥٤ مرة، فإنه نظرًا إلى أن كلمة "能力" وردت بمفردها داخل الذخيرة عدد ١١٧، أما كلمة "大" فقد وردت بمفردها في الذخيرة عدد ١٩٩١٣، ووفقًا للمعادلة رقم (٤-٣) فإن قيمة درجة التصاحب بين كل من "能力，弱" قد تجاوزت على العكس من ذلك نظيرتها بين كل من"能力，大". وهذه النتيجة نقطة تُحسَب لصالح مبدأ إلزامية التصاحب بين الكلمات.

وبالطريقة نفسها يمكن التوصل إلى النتائج التالية:

s（能力，强）=7.45

s（能力，差）=6.63

s（能力，小）=0.74

وحسب درجة قوة التصاحب يمكن ترتيب التصاحبات تنازليًا كما يلي:

s（能力，强）> s（能力，弱）> s（能力，大）> s（能力，小）

وهذا يوضح أن إمكانية وجود تصاحب فيما بين هذه المجموعات تتناقص تدريجيًا. ونظرًا إلى تقارب درجة قوة التصاحب وارتفاعها فيما بين المجموعات "能力，强"، و"能力，弱"، و"能力，差" يمكن اعتبارها تصاحبات لغوية. وعلى الرغم من أن نتيجة التصاحب لـ s（能力，大） منخفضة، فيمكننا عمل استقراء إضافي لهذه المجموعة لتقرير ما إذا كانت تُمَثِّل تصاحبًا لغويًا أم لا، أما نتيجة التصاحب s（能力，小） فتقترب من الصفر؛ ولذلك فمن الواضح أنها لا تمثل تصاحبًا لغويًا.

درجة انتشار التصاحب

ونظرًا إلى أن الكلمتين المتصاحبتين غالبًا ما يكون بينهما علاقة تركيبية مُحَدَّدَة، لذلك فإن فرصة ورود الكلمة w_i مع الكلمة w في موضع أو عدة مواضع بالنسبة إليها تكون كبيرة إلى حد بعيد؛ ومن ثَمَّ فإن قيمة الموضع j في الصيغة $r_j(w, w_i)$ تحدث بها قفزات كبيرة إلى حد بعيد. أما بالنسبة إلى الصيغ غير التصاحبية فنجد أن هذه القيمة لا تحدث بها فروق

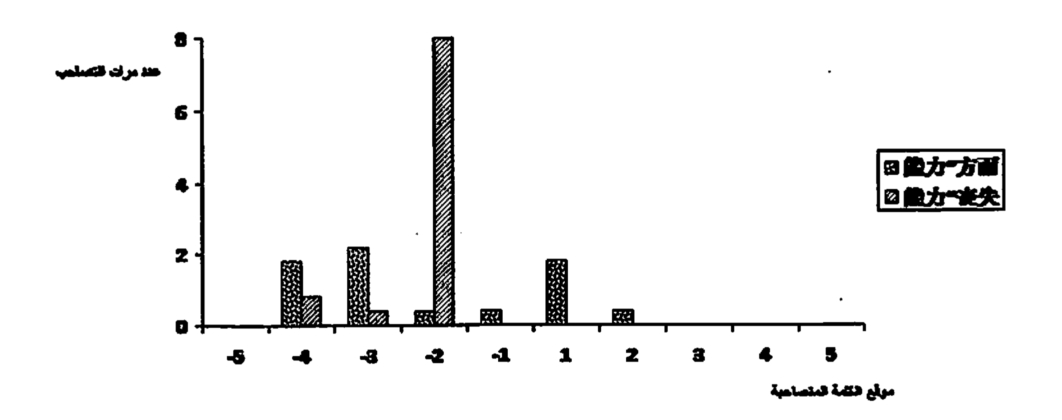
لافتة للنظر. ويُظهِر الشكل ٤-١ درجة توزيع التصاحب بين المجموعتين "能力，方面"، و"能力，丧失": فالمجموعة الأولى درجة توزيع التصاحب فيها شديدة التباين، أما المجموعة الثانية فدرجة توزيع التصاحب فيها متجانسة. والبيانات الإحصائية للمجموعتين كما يلي:

المجموعة الأولى: "能力，丧失"

r_{-4}（能力，丧失）$=r_{-3}$（能力，丧失）$=1$，r_{-2}（能力，丧失）$=8$，r_j（能力，丧失）$=0$ $(j=-5,-1,1,2,3,4,5)$

المجموعة الثانية: "能力，方面"

r_{-4}(能力，方面)$=r_1$(能力，方面)$=2$, r_{-3}(能力，方面)$=3$, r_{-2}(能力，方面)$=r_{-1}$(能力，方面)$=r_2$(能力，方面)$=1$, r_j(能力，方面)$=0$ $(j=-5,3,4,5)$



شكل (٤-١): توزيع التصاحب بين كلمتي المجموعتين

تساوت عدد مرات مشاهدات مجموعتي التصاحب "能力，丧失"، و"能力，方面" في الذخيرة اللغوية؛ حيث سجلت كل مجموعة ١٠ مرات ظهور. إلا أن الفرق بين توزيع التصاحب فيما بينهما كان كبيرًا. فبالنسبة إلى درجة الانتشار $u(w,w_1)$. للتوزيع $r_j(w,w_1)$ يمكن استخدام المعادلة التالية لحسابها:

$$u(w,w_1)=\frac{\sum_{j=-5}^{+5}[r_j(w,w_i)-\bar{r}(w,w_i)]^2}{10}$$

معادلة رقم (٤-٤)

حيث تشير $\bar{r}(w,w_1)$ إلى متوسط عدد مرات ظهور الكلمة w_i في المواضع المختلفة من الكلمة w.

$$\bar{r}(w,w_i)=\frac{\sum_{j=-5}^{+5}r_j(w,w_i)}{10}$$

معادلة رقم (٤-٥)

يُعَبِّر البسط في المعادلة رقم (٤-٤) عن درجة التفاوت والانسجام بين قيم الصيغة $r_j(w,w_i)$.

وعلى هذا الأساس يتم حساب درجة انتشار توزيع التصاحب بين المجموعتين كما يلي:

$\bar{r}$(能力，丧失)= (1+1+8) /10=1

u(能力，丧失)=

$((1-1)^2+(1-1)^2+(8-1)^2+7\times(0-1)^2)/10=5.60$

$\bar{r}$(能力，方面)= (2+3+1+1+2+1+4×0) /10=1

u(能力，方面)=

$((2-1)^2+(3-1)^2+(8-1)^2+3\times(1-1)^2+(2-1)^2$

$+4\times(0-1)^2/10=1.0$

عندما يكون مدى التغير في التوزيع كبيرًا، فمن الممكن أن يحدث تبعًا لذلك طفرة في أحد مواضع التصاحب، على سبيل المثال في الشكل ٤-١ عندما يكون موضع التوزيع $j=-2$ للتصاحب "能力，丧失" تكون r_{-2}(能力，丧失)=٨، وهذه القيمة تُمَثِّل تجمعًا لمشاهدات التصاحب عند هذا الموضع. عندما نحصل على قيمة $Z_j(w,w_i)$ في تجربة- Z بمعلومية $r_j(w,w_i)$ كما يلي:

$$Z_j(w,w_i)=\frac{r_j(w,w_i)-\bar{r}(w,w_i)}{\sqrt{u(w,w_i)}}$$

معادلة رقم (٤-٦)

يكون شرط حدوث طفرة لظهور التصاحب في الموضع j أن تكون $Z_j(w,w_i)$ كبيرة بقدر كاف.

طبقًا للشكل ٤-١، عندما يكون موضع التصاحب "能力，丧失"

$j=-2$ يتم التعويض في المعادلة كما يلي:

$$z_{-1}(能力，丧失)=\frac{8-1}{\sqrt{5.6}}=2.96$$

وتظهر نتيجة المعادلة أن قيمة r_{-2}(能力，丧失) أعلى من القيمة المتوسطة $\bar{r}$(能力，丧失) بمقدار ١. ٩٦ درجة انحراف معياري، مُمَثِّلَة بذلك طفرة في عملية التصاحب.

وقد قام سوين ماو سونغ بتصميم طريقة أكثر دقة في حساب طفرات التصاحب بالنسبة إلى اللغة الصينية كما يلي:

$is-peak(w,w_i)$

المدخلات: عدد مرات ظهور التصاحب بين أي زوج من التصاحبات w,w_i في كل موضع من مواضع التصاحب كما يلي:

$r_j(w,w_i)(j=-5,...,5)$

المخرجات: هل يوجد طفرة تصاحب أم لا؟ وما موقعها؟

يتم حساب متوسط عدد مرات التصاحب بين الكلمتين $\bar{r}(w,w_i)$ بالإضافة إلى نتيجة معادلة $z_j(w,w_i)(j=-5,...,5)$ بالنسبة إلى كل موقع من مواقع التصاحب.

تنفيذ الخطوات التالية مع كل حالات j:

إذا كانت $0.30 \le \bar{r}(w,w_i) < 1.00$ بالإضافة إلى أن:

$Z_j(w,w_i) \ge 2.50$ ، أو،

$1.00 \le \bar{r}(w,w_i) < 5.00$ بالإضافة إلى أن:

$Z_j(w,w_i) \ge 2.00$ ، أو،

$5.00 \le \bar{r}(w,w_i) < 10.00$ بالإضافة إلى أن:

$Z_j(w,w_i) \ge 1.50$ ، أو،

$\bar{r}(w,w_i) \ge 10.00,$ ، بالإضافة إلى أن $Z_j(w,w_i) \ge 1.00$

تكون j موضع طفرة في التصاحب. وإلا فإن j لا تمثل طفرة تصاحبية.

طريقة الحساب السابقة تُقَسِّم متوسط التصاحبات بين الكلمتين $\bar{r}(w,w_i)$ إلى عدة قطاعات، وتقوم بحساب قيمة $Z_j(w,w_i)$ لكل قطاع يُمَثِّل طفرة تصاحبية. والأرقام الموضَّحَة سالفًا قد تم تحديدها من خلال التجارب، وبصورة عامة فإن قيمة متوسط المشاهدات $\bar{r}(w,w_i)$ قليلة لأن عدد مرات مشاهدة التصاحب منخفضة، وعلى ذلك تكون درجة مرجعية البيانات الإحصائية ضعيفة؛ ولذلك ينبغي تعديل رقم البداية إلى رقم أعلى من ذلك؛ وعلى العكس من ذلك عندما تكون البيانات الإحصائية وافية، يمكننا تعديل رقم البداية إلى قيمة أقل. على سبيل المثال نجد أن التصاحب "能力， 丧失"

ظهر في ذخيرة XH-CORPUS عشر مرات، وبناءً على ذلك كانت قيمة متوسط عدد مرات التصاحب ١. ٠٠=(丧失،能力)$\bar{r}$، وبذلك تكون درجة ضجيج البيانات مرتفعة إلى حد كبير، ولو أننا قمنا بتعديل قيمة الفرض المبدئي إلى رقم أكبر (أكبر من ٢. ٠٠) فسوف تكون النتيجة أن ٢. ٩٦=(丧失،能力)Z_{-2} وهذا الرقم يمكن أن يُمَثِّل مرجعية لحساب درجة ظهور الطفرة التصاحبية عندما تكون $j=-2$. أما التصاحب "能力，大" فقد ظهر في الذخيرة عدد ٥٤ مرة، وعلى ذلك يكون مقدار متوسط عدد مرات الظهور 5.40=(能力,大)$\bar{r}$، وهذا الرقم يمكن الاعتماد عليه، وعلى هذا يمكننا تقليل درجة البداية قليلاً (أكبر من ١. ٥)، وبذلك تكون ١ ٨٤=(大،能力)Z_1 وهذا الرقم يمكن أن يُعتَبَر مرجعية لتقدير ظهور الطفرة التصاحبية عند $j=1$.

إن قيمة متغير درجة التشتت والطفرة التصاحبية من شأنهما أن يَقدّما بيانات مرجعية تُفِيد في عمل دراسات كمية عن أبنية التصاحب اللغوي. ويرى سون ماو سونغ أنه على الرغم من أن هذين المتغيرين لهما قيمة مرجعية عالية، فهما لا يكفيان لإثبات وجود التصاحب بمفردهما.

وهناك عدد من التصاحبات تكفي درجة قوة التصاحب بين كلماتها لإثبات التصاحب، ولا يستلزم ذلك وجود درجة انتشار عالية، وأكثر من ذلك قد لا يحدث بينها طفرة تصاحبية (من الواضح أننا لو انطلقنا فقط من درجة توزيع التصاحب فإن الطفرة التصاحبية سيكون مطلبًا صعبَ المنال بالنسبة إلى الكلمات المتصاحبة). ومن هنا، فإن الاحتكام إلى متغيّرَي درجة الانتشار

والطفرات التصاحبية لا يتم اللجوء إليه إلا عندما تكون المعلومات الخاصة بقوة التصاحب غير كافية لاتخاذ قرارات صائبة.

وفيما يلي نعرض مجموعة المتواليات المنطقية التي صممها سوين ما سونغ للحكم على وجود علاقة تصاحبية بين كلمتين، التي اعتمد فيها على حساب المؤشرات الثلاثة: درجة قوة التصاحب (معادلة ٤-٣)، ودرجة الانتشار (معادلة ٤-٤)، والطفرة التصاحبية (معادلة ٤-٦)،

وقد أطلق سوين ماو سونغ على تلك المعادلات اسم $is-collocation(w, w_i)$

المدخلات: قوة التصاحب $s(w, w_i)$، ودرجة انتشار التصاحب $u(w, w_i)$، والمتوسط الحسابي لعدد مرات التصاحب $\bar{r}(w, w_i)$، بالإضافة إلى عدد مرات التصاحب في كل موقع للتصاحب $Z_j(w, w_i)(j=-5,...,5)$ لأي زوج من الكلمات w, w_i.

المخرجات: الحكم على وجود علاقة تصاحبية بين كل من w, w_i من عدمه.

إذا كانت $\bar{r}(w, w_i) < 0.30$، فإن w, w_i لا ترتبطان بعلاقة تصاحبية.

إذا كانت $s(w, w_i) \geq 4.50$، إذن فإن w, w_i ترتبطان بعلاقة تصاحبية.

وإلا، إذا كانت $3.50 \leq s(w, w_i) < 4.50$، بالإضافة إلى أن $u(w, w_i) \geq 10.00$

إذن فإن w, w_i ترتبطان بعلاقة تصاحبية.

وإلا إذا كانـــــت $2.50 \leq s(w, w_i) < 3.50$ ، بالإضـــــافة إلـــــى أن $u(w, w_i) \geq 20.00$

يتم الحكم على أن w, w_i ترتبطان بعلاقة تصاحب.

وإلا إذا كانت $s(w, w_i) \geq 2.00$

يتم التعديل إلى معادلة الطفرة التصاحبية $is - peak(w, w_i)$.

إذا ظهرت طفرة تصاحبية

إذن فإن w, w_i ترتبطان بعلاقة تصاحب.

وإلا فإن w, w_i لا ترتبطان بعلاقة تصاحب.

من خلال المتواليات المنطقية السابقة يمكننا أن نقرر أن هناك ثلاثـــة شروط تتأكد عندها العلاقة التصاحبية بين كلمتين:

(١) عندما تكون درجة القوة التصاحبية بين الكلمتين كبيـــرة بدرجـــة كافية، فلا حاجة لحساب درجة الانتشار.

(٢) كلما تناقصت درجة القوة التصاحبية زادت الحاجة إلـــى حـــساب درجة الانتشار.

(٣) عندما تنخفض درجة القوة التصاحبية إلى حد معين، يتأكد حدوث طفرة تصاحبية.

وهناك ثلاثة شروط لنفي وجود التصاحب:

(١) أن يكون عدد مرات ظهور الكلمتين معًا منخفضًا بدرجة كبيرة، والبيانات المُحَصَّلة ليس لها مغزى إحصائي.

(٢) أن تكون درجة القوة التصاحبية منخفضة، ولم يصاحبها طفرة تصاحبية.

(٣) أن تكون درجة القوة التصاحبية منخفضة وعلى الرغم من حساب درجة التشتت والطفرة التصاحبية، فإن النتائج النهائية ليست ذات قيمة تُذكَر.

وقد قام سوين ماو سونغ بعمل تحليل تفصيلي لجميع أنواع حالات الظهور المشترك بين كلمة "能力" والكلمات الأخرى في محيط قدره ±5 كلمات وذلك على ذخيرة **XH-CORPUS** التي يصل حَجْمُها إلى سبعة ملايين ومائة ألف كلمة. وكانت نتيجة التجربة ما يلي: وصل عدد مرات ظهور كلمة "能力" داخل الذخيرة إلى ٢٢٤١ مرة، (أي أن $r(w = 2241, w = \text{能力})$) كما وصل عدد الكلمات التي اقترن ظهورها مع كلمة "能力" في سياق يصل طوله إلى ±5 كلمات إلى ١٩٣٢ كلمة، وقد تم التعامل مع جميع هذه الكلمات على أنها مُرَشَّحَة لأن تَكُون من مصاحبات كلمة "能力". تم استبعاد عدد ١٣١٧ كلمة بعد تحقيق الشروط الثلاثة لنفي وجود التصاحب: وكانت تفاصيل قرار الاستبعاد كما يلي:

(١) تم استبعاد عدد ٩٦٢ كلمة ينطبق عليها الشرط الأول لنفي التصاحب.

(٢) تم استبعاد عدد ٢٠١ كلمة ينطبق عليها الشرط الثاني لنفي التصاحب.

(٣) تم استبعاد عدد ١٥٤ كلمة ينطبق عليها الشرط الثاني لنفي التصاحب. تم الإبقاء على عدد ٦١٥ كلمة بعد تحقيق الشروط الثلاثة لإقرار التصاحب:

أ - تم تأكيد علاقة التصاحب لعدد ٤١١ كلمة وفقًا للشرط الأول لتأكيد علاقة التصاحب.

ب - تم تأكيد علاقة التصاحب لعدد ٢٧ كلمة وفقًا للشرط الثاني لتأكيد التصاحب.

جـ - تم تأكيد علاقة التصاحب لعدد ١٧٧ كلمة وفقًا للشرط الثالث لتأكيد التصاحب. وقد كان من بين الكلمات التي تم تأكيد علاقة التصاحب معها عدد ٨٨ كلمة، ولا يمكن وجود علاقة تصاحب بينها وبين كلمة "能力" (مثل كلمات الأعداد"千，一" بمعنى "واحد، ألف")، والكلمات المساعدة (مثل"了"的，")، وأداوت العطف (مثل "和，无论" بمعنى "و، سواء")، والظروف (مثل"不，较" بمعنى "لا النافية، إلى حد ما")، وما إلى ذلك، وباستخدام برمجية بسيطة من برمجيات الترشيح يمكن استبعاد تلك الكلمات من النتائج النهائية. وبالإضافة إلى ذلك، فقد كان هناك عدد من أخطاء الحكم على وجود علاقة تصاحب وصلت إلى ٢٩ حالة؛ بسبب وجود خطأ في التقسيم الآلي لحدود الكلمات الصينية، على سبيل المثال عدم اشتمال معجم التقسيم الآلي لحدود الكلمات لكلمة "调控" بمعنى "يتحكم"، الأمر الذي أدى ببرمجية التقسيم إلى اعتبار أن "调控能力" بمعنى "القدرة على التحكم" يتم تقسيمها إلى "调/控/能力" ومن هنا فقد أدى ذلك إلى أن برمجية الحُكم على

التصاحبات اللغوية اعتَبَرَت أن كلا من "条，能力"، و"控，能力" تمثلان تصاحبين لغويين بمعنى "القدرة على النقل"، و"القدرة على توجيه الاتهام" على التوالي وهذا خطأ بالطبع. والحقيقة أن التصاحب هو "调控，能力". وبعد استبعاد عنصري الخطأ السابقين كانت حالات التصاحب التي تم تقريرها آليًّا هي ٤٩٨ حالة تصاحب. وبعد إجراء مراجعة من خلال العنصر البشري، تم إقرار عدد ١٦٩ حالة تمثل تصاحبًا حقيقيًّا. وبعبارة أخرى، فإن نسبة دقة الحاسب في استخراج التصاحبات اللغوية وتحديدها قد وصلت إلى ١٦٩/٤٩٨=%٣٣,٩٤. ويعرض لنا الجدول ٤-٥ جزءًا من نتيجة التجربة التي تم إجراؤها على كلمة "能力" في اللغة الصينية.

جدول ٤-٥ عرض جزء من البيانات التجريبيةلـ

($w=$能力 $r(w)=2241$)

م	w_i	$r(w_i)$	$r(w,w_i)$	$s(w,w_i)$	$u(w,w_i)$	موضع الطفرة التصاحبية	الحكم الآلي عما إذا كان هناك تصاحب أم لا	الحكم الآلي عن صحة الحكم	نوع التصاحب من حيث التركيب
1	吞吐	78	78	11.30	547.56	1-	نعم (تأكيد 1)	صحيح	موصوف وصفة (*)
2	分辨	25	3	8.57	0.81	1-	نعم (تأكيد 1)	صحيح	موصوف وصفة
3	强	1651	91	7.45	138.29	1	نعم (تأكيد 1)	صحيح	مسند إليه ومسند
4	弱	177	9	7.33	2.69		نعم (تأكيد 1)	صحيح	مسند إليه ومسند
5	提高	6058	205	6.75	187.25	4-،3-	نعم (تأكيد 1)	صحيح	مسند إليه ومسند
6	查	638	200	6.63	18.20	1	نعم (تأكيد 1)	صحيح	مسند إليه ومسند- فعل ومفعول
7	具有	2749	51	5.88	45.49	3-،2-	نعم (تأكيد 1)	صحيح	فعل ومفعول
8	和	15377	278	5.84	632.36	3-،2-،1	نعم (تأكيد 1)	خطأ	
9	石油	1641	18	5.12	18.20	2-	نعم (تأكيد 1)	خطأ	
10	有	30354	212	4.47	5788.36	1-	نعم (تأكيد 1)	صحيح	فعل ومفعول
11	使	8701	62	4.49	57.30	5-،4-	نعم (تأكيد 2)	خطأ	
12	组织	7760	24	3.29	22.64	1-	نعم (تأكيد 2)	صحيح	موصوف وصفة
13	问题	12476	28	2.83	24.16	2-	نعم (تأكيد 2)	خطأ	
14	拥有	1141	5	3.80	0.85	3-	نعم (تأكيد 3)	صحيح	فعل ومفعول
15	大	19913	54	3.10	3.84	1	نعم (تأكيد 3)	صحيح	مسند إليه ومسند

تابع الجدول السابق

م	w_i	$r(w_i)$	$r(w,w_i)$	$s(w,w_i)$	$u(w,w_i)$	موضع الطفرة التصاحبية	الحكم الآلي عما إذا كان هناك تصاحب أم لا	الحكم الآلي عن صحة الحكم	نوع التصاحب من حيث التركيب
16	不	17409	37	2.75	19.21	1	نعم (تأكيد 3)	خطأ	
17	而	6908	17	2.96	3.01		لا (رفض 2)	صحيح	
18	国家	18986	38	2.67	13.16		لا (رفض 2)	صحيح	
19	接受	1729	6	3.46	0.44		لا (رفض 2)	خطأ	موصوف وصفة
20	要	15404	18	1.89	5.36	3-	لا (رفض 3)	صحيح	
21	小	9481	5	0.74	0.85	3	لا (رفض 3)	صحيح	
22	活动	7428	8	1.77	2.36	1-	لا (رفض 3)	خطأ	موصوف وصفة
23	民族	5473	2	0.21	0.16		لا (رفض 1)	صحيح	
24	东	5471	1	0.78-	0.09		لا (رفض 1)	صحيح	
25	涵蓄	1	1	11.63	0.09		لا (رفض 1)	خطأ	موصوف وصفة

وقد عقد سوين ماو سونغ مقارنةً بين كل من نتائج التصاحبات التي توصل إليها بعد مراجعتها من خلال العنصر البشري وتصاحبات كلمة "能力" التي وردت في (المعجم الكبير في اللغة الصينية الحديثة) "现代汉语辞海"، نظرًا إلى أن هذا المعجم هو أكثر معجم قام بتغطية الخصائص التصاحبية للكلمات الحقيقية في اللغة الصينية الحديثة بصورة وافية. ويعرض القسم (a) من الجدول رقم ٤-٦ عناصر التصاحب التي تم تكرارها بين نتيجة التجربة والمعجم، أما الجزء (b) من الجدول فيعرض التصاحبات التي لم يذكرها المعجم. ويمثل هذا الجدول في الوقت نفسه القائمة الكلية للتصاحبات اللغوية مع كلمة "能力" التي تم استخراجها آليًّا بعد مراجعتها من خلال العنصر البشري.

جدول ٤-٦ التصاحبات اللغوية التي تم استخراجها آليًا بعد مراجعتها عن طريق عنصر بشري

(المجموع ١٦٨ تصاحبًا)

(a)

培养 判断 鉴赏 生产 竞争 制造 运输 加工 支付 偿还 平衡 消化 吸收 繁殖 实际 具备 缺乏 提高 强弱 大差 有限 增强 具有 业务 劳动 适应 领导 组织 分析 保护 发挥 工作 技术 专业 管理 创造 运输 发电 丧失 防御 指挥

(b)

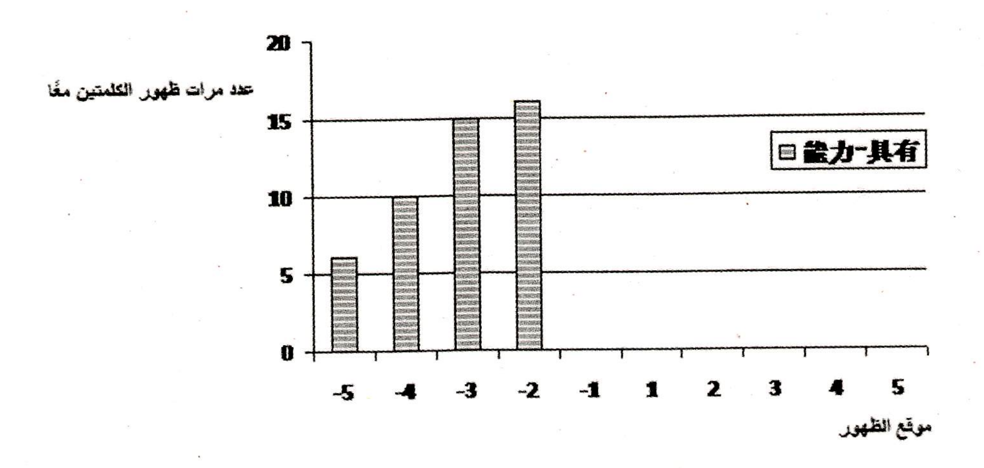
反应 扩大 综合 形成 达到 设计 抗灾 开采 影响 排水 客运 保障 承受 一定 执政 反应 安置 配套 不足 超过 出口 自立 创汇 动手 吞吐 增加 运行 足够 防务 操作 处理 作战 通信 同等 自给 自理 防守 减弱 现有 约束 作业 防卫 鉴别 通航 负重 不够 生存 隐蔽 科研 失去 抗病 炼油 腐蚀 后续 识别 抗旱 削弱 限制 识字 存储 自主 对抗 核算 机动 消费 分流 超出 防洪 自卫 干扰 免疫 再生 信任 过剩 供给 应急 饲养 运算 扑救 防疫 驾驭 筛选 参政 相应 采油 整体 通行 核定 载荷 维修 运载 接待 保存 分辨 保鲜 装备 耐寒 通车 转换 防范 自救 联运 决策 独到 起重 输送 新有 开发 服务 群众 发展 测量 显示 突破 依靠 强化 控制 经营 供应 下降 监督 低核 拥有

ويشير الاستقراء الذي أجراه سوين ماو سونغ على التصاحب داخل الذخيرة إلى ما يلي:

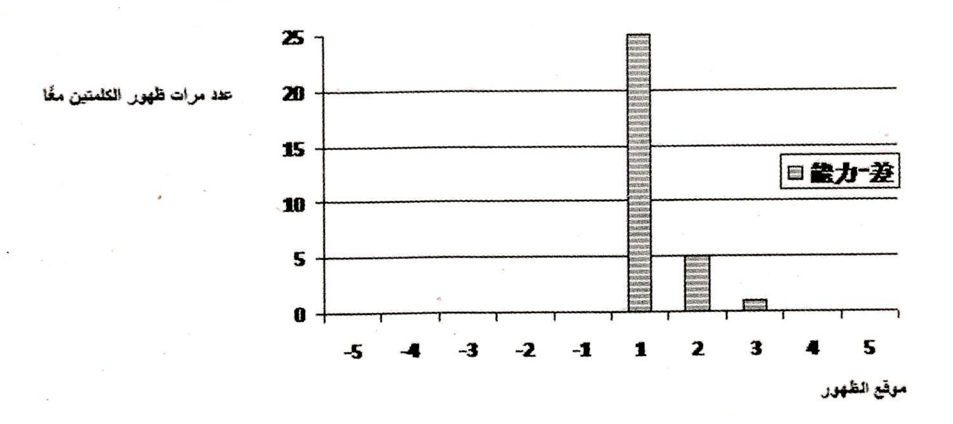
(١) أن كلا من القيم الإحصائية الثلاث: القوة التصاحبية $s(w, w_i)$، ودرجة الانتشار $u(w, w_i)$ والطفرة التصاحبية قيم إحصائية مناسبة لعمل تحليل كمي للتصاحب اللغوي، إلا أنها لا تتعدى كونها مؤشرات نسبية، والدليل على ذلك أن الجدول رقم ٤-٥ قد احتوى أمثلة تُمَثِّل خطئًا في الحكم؛ نظرًا إلى عوامل مختلفة.

(٢) أن الخصائص التوزيعية للبيانات الإحصائية تعكس بدرجة كبيرة الخصائص التركيبية للتصاحب.

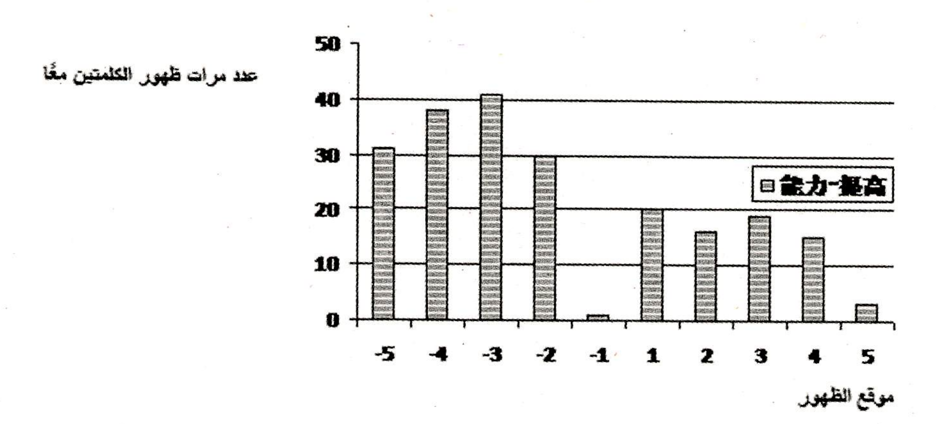
فالشكل ٤-٢ يوضح أن جميع حالات التصاحب بين كل من "能力，具有" عبارة عن تركيب فعل ومفعول (حيث ظهرت الطفرة التصاحبية في الموقعين -٢،-٣)، أما الشكل ٤-٣ فيُظْهِر أن جميع حالات التصاحب بين كل من "能力，差" عبارة عن تركيب إسنادي (حيث ظهرت الطفرة التصاحبية في الموقع +١)، أما الشكل (٤-٤) فيُظْهِر أن التصاحب "能力，提高" قد يكون تركيبًا من فعل ومفعول (حيث ظهرت الطفرة التصاحبية في الموقعين -٣،-٤)، وقد يكون أيضًا تركيبًا إسناديًّا؛ أما الشكل ٤-٥ فيُظْهِر أن التصاحب بين "能力，吞吐" قد حدث معه طفرة تصاحبية كبيرة (في الموقع -١)، وهذا يُعَبِّر عن نموذج من نماذج التصاحب اللغوي القياسي في اللغة الصينية وهو التعبيرة الاسمية المكوَّنة على هيئة مضاف ومضاف إليه أو صفة وموصوف.



شكل ٤-٢ توزيع ظهور "能力 - 具有"

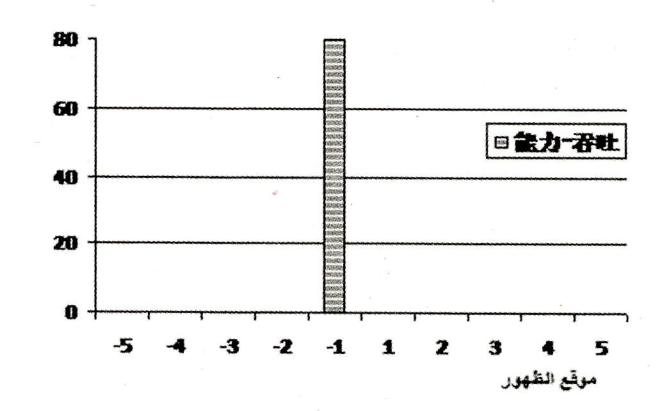


شكل ٤-٣ توزيع ظهور "能力 - 差"



شكل ٤-٤ توزيع ظهور "能力 -提高"

عدد مرات ظهور الكلمتين معًا



الشكل ٤-٥ توزيع "能力-吞吐"

(٣) أن التصاحب يتأثر بتغير المجال اللغوي. فالتصاحبين اللغويين المتفق عليهما من الجميع "读能力" بمعنى "القدرة على القراءة"، و"写作能力" بمعنى "القدرة على الإنشاء" لم يظهرا في ذخيرة XH-CORPUS بسبب اختلاف المجال اللغوي.وحتى إذا تم تحديد المجال اللغوي، فإن حجم الذخيرة وطرق اختيار عينات النصوص من شأنه أن يؤثر تأثيرًا شديدًا في عملية إحصاء التصاحب، كما في الجدول ٤-٥ حيث نجد أن التصاحبات "接受能力" (درجة القوة التصاحبية ٣. ٤٦)، و"活动能力" (الطفرة التصاحبية عند الموقع -١)، و"涵蓄能力" (درجة القوة التصاحبية ١١. ٦٣)، على الرغم من أن عدد مرات الظهور كانت مرة واحدة، فالجميع يتفق على أن كل هذه المجموعات تمثل تصاحبات لغوية. إلا أنه بالنظر إلى البيانات الإحصائية الخاصة بها نجد أنها بمشقة قد رُشحت لاختبار التصاحب؛ ولكن نظرًا إلى انخفاض عدد مرات ظهورها داخل ذخيرة XH-CORPUS فقد تم استبعادها من القائمة النهائية للمتصاحبات.

(٤) أنه بالنسبة إلى قيمة الفرض المبدئي يتم تحديدها كقيمة وسط بين كل من معدل الدقة (يُقصد به نسبة ما تم إقراره من تصاحبات عن طريق تدخل العنصر البشري مقارنة بما تم اكتشافه بواسطة الآلة) ومعدل الارتداد (يقصد به نسبة التصاحبات التي أقرتها الآلة مقارنة بالعدد الكلي للتصاحبات المُرَشّحة داخل الذخيرة). وبصورة عامة،

كلما ارتفعت قيمة البداية، ارتفع تبعًا لذلك معدل الدقة، وعلى العكس من ذلك، كلما انخفضت قيمة البداية، ارتفع معدل الارتداد وانخفض تبعًا لذلك معدل الدقة.

وقد كانت الدراسة الاستقرائية لسوين ماو سونغ على الذخيرة تسعى إلى اكتشاف أكبر قدر من التصاحب اللغوي بغض النظر عن دقة النتائج النهائية، على اعتبار أن زيادة العدد أفضل من القصور في الحصر؛ ولذلك فقد كان متحفظًا في تحديد قيمة البداية.

وإذا أردنا أن نعرف هل ٣٠% تقريبًا كمعدل دقة لنتائج التصاحبات فيما يتعلق بكلمة "能力" منخفض أم لا، يمكننا أن نطلع على ما أشار إليه سمادجا من أن معدل دقة نتائج التصاحبات التي تم اتباعها يدويًا في تأليف معجم **Oxford English Dictionary (OED)** كانت ٤% تقريبًا. وعن طريق المقارنة يمكننا القول: إن الاستعانة بالحاسب الآلي في اكتشاف التصاحب اللغوي يفوق العمل اليدوي، وأن فاعلية النتائج باستخدام الحاسوب أعلى بدرجة كبيرة. بالإضافة إلى ذلك فإن استخدام الأسلوب اليدوي من شأنه أن يتأثر سلبًا بالعوامل البشرية؛ لأن الإحساس اللغوي يختلف من باحث إلى آخر؛ لأن ذلك له علاقة باختلاف المستوى اللغوي والخلفية المعرفية لكل شخص، ومن هنا تكون أحكام البشر على التصاحب خالية من الموضوعية، ومن الصعب التنسيق فيما بينهم. أما استخدام الحاسب الآلي في عمل تحليل كمي للتصاحبات اللغوية داخل ذخيرة لغوية كبيرة الحجم من المؤكد أنه يخفف قوة ضغط العمل على اللغويين، ويرفع من كفاءة النتائج ومساحة التغطية.

الفصل الرابع

دراسات التصنيف الدلالي للأسماء في اللغة الصينية

من المعلوم لدى الجميع أن التصاحب بين الاسم والكلمة الكمية(١) من الظواهر المهمة التي تميز اللغة الصينية عن اللغات الهند أوروبية. ويرى العديد من العلماء أن الفائدة الأساسية لتراكيب الاسم + الكلمة الكمية في اللغة الصينية الحديثة هي تصنيف الأسماء دلاليًا. ولذلك فإن إجراء استخراج كمي على ذخيرة كبيرة الحجم للتعرف على البيانات الكَمِّيَة لتصاحبات الاسم + الكلمة الكمية من شأنه أن يساعد في وضع آلية للتصنيف الدلالي لأسماء اللغة الصينية[88]. وكانت تلك هي الخلاصة التي توصل إليها العلماء التايوانيون خوانغ جو رن، وتشين كه جيان، وجاو مينغ من خلال الدراسة التي أشرفوا عليها. وقد استعانت هذه الدراسة بشكل مباشر بالبيانات التصاحبية عن الأسماء + الكلمات الكمية التي استخرجها مؤلف الكتاب من ذخيرة أبحاث اللغة الصينية. وعلى أساس تلك البيانات الكمية الثابتة استعانوا بمعادلة الضجيج المعلوماتي في حساب محتوى المعلومات داخل تركيب

---

(١) انظر الفصل السابق صفحة ٢٤٤. (المترجم)

الاسم + الكلمة الكمية مع الاستعانة بأسلوب المُتَّجَهات[1] (vectors) في حساب مقدار التقارُب (الدلالي) بين مجموعتين من المجموعات الاسمية. وقد توصل البحث إلى أن استخدام الطريقة سالفة الذكر من شأنه أن يُرشِد في التوصل إلى نظام يمكن الاعتماد عليه بشكل كبير في تصنيف الأسماء في اللغة الصينية. وقد استعان الكِتاب الذي بين يدي القارئ بهذا المثال للتعريف بالدراسات اللغوية القائمة على ذخائر لغوية، ويرجع ذلك إلى السببين التاليين:

(١) إن هذه الدراسة قد اعتمدت على أمثلة حقيقية عن تراكيب الاسم + الكلمة الكمية والبيانات الإحصائية المتعلقة بها، التي تم استخراجها مباشرة من ذخيرة معهد الدراسات المركزي بتايوان، وقد عُرِضَت نتيجة الدراسة في (معجم الكلمات الكمية المستخدمة في تايوان) الصادر عام ١٩٩٧ عن درا نشر الصحافة بتايوان.

(٢) إن عمل تصنيف لأسماء اللغة الصينية، وفقًا لحالات اقترانها مع الكلمات الكمية يتحتم عليه إجراء تمثيل شكلي مستقل عن حالات اقتران الكلمات الكمية مع الأسماء، بالإضافة إلى عرض المعادلات الخاصة بحساب درجة التشابه فيما بينها. ففي هذه الدراسة، يتم

---

(١) المُتَّجه (vector) مصطلح في الفيزياء يعني قيمة لها حجم واتجاه، ويتم تمثيلها بسهم يُعَبِّر عن اتجاهها من نقطة إلى أخرى، والكميات المُتَّجَهة تختلف عن الكميات العددية في أنه لا يمكن الاكتفاء عند تحديدها بقيمة عددية فقط، وإنما نحتاج أيضًا إلى تحديد اتجاهها. وقد استعان علماء اللغة بهذا المفهوم في حساب درجة اتجاه التصاحب بين الكلمات كما يشير النص. (المترجم)

حساب الحمل المعلوماتي[1] الخاص بالكلمة الكمية (يطلق عليه أيضًا اسم الإنتروبي أو مقياس درجة الفوضى) من خلال عدد الأسماء التي يمكن أن تقترن معه؛ أما معاني كل اسم من تلك الأسماء فيتم التعبير عنها من خلال مجموعة الكلمات الكمية التي يمكن أن يَقْبَل الاقتران بها (المتَّجَهات التي تتكون من ذلك الاقتران). ولذلك فمن خلال حساب الفرق بين مُتَّجَهات الأسماء يمكننا عمل تصنيف لمجموعة الأسماء داخل اللغة؛ ومن ثَمَّ التوصل إلى آلية للتصنيف الدلالي لها. إن التمثيل الشكلي للمادة اللغوية التي يتم معالجتها وتصميم معادلة مناسبة لحساب طريقة التعامل معها تمثل أسلوبًا مثاليًا للنمذجة اللغوية (modeling) بالإضافة إلى إمكانية تعميمه.

وحتى نفهم العمل البحثي الذي قام به خوانغ جو رن وفريقُه البحثي، من الضروري أن نشرح بشيء من التفصيل مفهوم الحمل المعلوماتي الخاص بحَدَث عشوائي والذي ورد في نظرية المعلومات للعالم كلود شانون **(Shannon)[89]**. بافتراض أن إس تُعَبِّر عن الأحداث العشوائية $E_1, E_2, ..., E_n$، فإن احتمال حدوث

(١) الحمل المعلوماتي أو (الإنتروبي) **(entropy)** أو مقياس درجة الفوضى مصطلح في الفيزياء والكيمياء يشير إلى التغير والتحول إلى حالة من الفوضى على المستوى الجزيئي في نظام ما، كمثال اختلاط نقطة حبر بالماء، في البداية يكون توزيع ذرات الحبر في الماء غير منتظم، وبعد انقضاء فترة زمنية يكتسب الماء لونا متجانسًا، ومن هنا يمكننا اعتبار درجة الحيرة أو الإنتروبي مقياسًا لعدم التأكد أو مقياسًا لعدم الانتظام. (المترجم)

أي حدث من تلك الأحداث E_i يُرْمَز له بالرمز p_i. وطبقًا للمبادئ الأساسية لعلم الاحتمال فإن: $0 \le p_i \le 1, p_1 + p_2 + \ldots + p_n = 1$.

والحمل المعلوماتي لحدث مُحَدَّد E_i يُحْسَب من المعادلة:

$$I(E_i) = -\log_2 p_i \quad \textbf{(bit)}$$

معادلة رقم (٤-٧)

وطبقًا لبديهيات علم الاحتمال يمكننا أن نَعْرِف أن الحمل المعلوماتي I أكبر من الصفر؛ وأنه كلما قَلَّت قيمة احتمال ظهور الحدث، زادت قيمة الحِمْل المعلوماتي المصاحب له. الأحداث المؤكَّدَة يكون الحمل المعلوماتي الخاص بها ($p_i = 1$) يساوي صفرًا؛ أي أن ظهورها من الأحداث المتوقعة، وعلى هذا فهي لا تحمل إلينا أية معلومات.

يمكن الاستعانة بقيمة درجة الحيرة (الإنتروبي) $H(S)$ في الرسالة الكلامية للتعبير عن قيمة المتوسط الإحصائي لكمية المعلومات الخاصة بكل عنصر من عناصر الرسالة الكلامية S؛ وذلك وفقًا لكمية المعلومات $I(E_i)$ التي يحملها كل حدث داخل الرسالة الكلامية (الحِمْل المعلوماتي) S، كما توضح المعادلة التالية:

$$H(S) = -\sum_{i=1}^{n} p_i \log_2 p_i \quad \textbf{(bit)}$$

معادلة رقم (٤-٨)

ونظرًا إلى أن قيمة الحِمْل المعلوماتي للحدث العشوائي تزداد بزيـــادة درجة عدم اليقين من حدوثه؛ لذلك يمكننا الاعتماد على قيمـــة الحيـرة فـــي الرسالة الكلامية (الإنتروبي) كمقياس لدرجة عدم اليقين فـــي حـــدوث تلـــك الرسالة. وعلى هذا الأساس، لا يمكن أن تكون درجة الحيرة فـــي الرســـالة الكلامية (الإنتروبي) قيمة سالبة، فكلما زادتِ قيمة درجة الحيرة ارتفعت تبعًا لها قيمة عدم اليقين في حدوث الرسالة. وعلى العكس من ذلك، إذا كانت S رسالة كلامية مؤكدة الحدوث، فإن قيمة درجة الحيرة لها تكون صفرًا.

على فرض أن هناك رسالة كلامية بها عدد N من الأسماء، ومن بين هذه الأسماء هناك الاسم n يمكن أن يتصاحب ظهوره مع الكلمة الكمية X، إذن الحِمْل المعلوماتي لهذه الكلمة الكمية يساوي الفرق بين درجتي الحيـــرة (الإنتروبي) التاليتين:

$$I(X) = H(N) - H(n)$$

معادلة رقم (٤-٩)

وإذا كان احتمال ظهور كل من الاسم والكلمة الكميـــة متـــساوٍ، فـــإن احتمال ظهور اقتران بين كل اسم ضمن مجموعة الأسماء N وكلمة كميـــة معينة يساوي $\frac{1}{N}$، ودرجة الحيرة له يتم حسابها كما يلي:

$$H(N) = -\sum_{i=1}^{n} \frac{1}{N} \log_2(\frac{1}{N}) = \log_2 N$$

معادلة رقم (٤-١٠)

وبالطريقة نفسها فإن درجة الحيرة لاقتران الاسم n داخل رسالة كلامية مع كلمة كمية معينة يتم حسابها من المعادلة التالية:

$$H(N) = \sum \frac{1}{n}\log_2(\frac{1}{n}) = \log_2 n$$

معادلة رقم (٤-١١)

ومن هنا فإن الحمل المعلوماتي للكلمة الكمية X يُحْسَب من المعادلة التالية:

$$I(x) = \log_2 N - \log_2 n$$

معادلة رقم (٤-١٢)

ونظرًا إلى أن N قيمة ثابتة، فإن المعادلة السابقة تُخْبِرُنا أنه كلما قَلَّ عدد الأسماء التي تقترن بكلمة كمية معينة، زاد الحِمْل المعلوماتي لهذه الكلمة الكمية، وزادت تبعًا لذلك قدرة هذه الكلمة في المساهمة في عمل تصنيف دلالي للأسماء في اللغة الصينية. وهذا يتفق إلى حدٍّ كبير مع الشعور البديهي لكل منا.

عندما نستخدم مُتَّجَهًا متعدد الأبعاد للتعبير عن كل اسم (أو مجموعة أسماء) داخل السلسلة الكلامية، فإن كل بُعْد من هذه الأبعاد يُعَبِّر عن كلمة كمية محددة داخل السلسلة الكلامية، وقيمة المُتَّجه في بُعْد معين من تلك الأبعاد يساوي الحِمْل الملعوماتي لهذه الكلمة الكمية (المعادلة ٤-١٠) وإذا كان هذا الاسم (أو مجموعة الأسماء) لا يمكن أن تظهر مع كلمة كمية معينة، فإن قيمة المُتَّجه الخاص بها يساوي صفرًا. وعلى أساس ما يُعَبِّر عنه

المُتَّجه، يمكن قياس درجة التقارب الدلالي (affinity) بين اسمين (أو مجموعتين من الأسماء) عن طريق حساب الفرق بين قيمة مُتَّجَهيهما.

وعلى أساس الفرض السابق، فإن فريق البحث اتبع طريقة لتصنيف الأسماء تقوم بإنشاء تصنيف جديد للأسماء عن طريق ضم أكثر اسمين متشابهين معًا في تصنيف واحد، وقد نشأ عن ذلك في النهاية شجرة دلالية للأسماء. وقد كانت خطوات الحساب التي أدت إلى عمل تصنيف للأسماء كما يلي:

(١) من خلال المعادلة (٤-١٢)، تم حساب الحِمْل المعلوماتي لعدد ١٨٢ كلمة كمية.

(٢) تم اعتبار أن كل كلمة كمية من الـ ١٨٢ كلمة يُمَثِّلها بعدٌ واحد داخل مُتَّجه ذي ١٨٢ بُعْدًا، الحِمْل المعلوماتي لكل كلمة كمية يساوي قيمة من قيم المُتَّجه متعدد الأبعاد.

(٣) كل نوع من أنواع الأسماء يساوي مُتَّجهًا من تلك المُتَّجهات، وهذا المُتَّجه يتم الاعتماد عليه في تحديد قيمة المُتَّجه الخاص بكل الكلمات الكمية التي تتصاحب مع هذا الاسم وفي وضع تعريف لهذا التصنيف من تصنيفات الأسماء.

(٤) بصورة متكررة، يتم ضم أي تصنيفين متجاورين من تصنيفات الأسماء بشكل عشوائي في مجموعات لتكوين تصنيف جديد، واستخدام قيمة المُتَّجه المقابل لهما في التعبير عن ذلك التصنيف،

وهذا المُتَّجه الجديد عبارة عن متوسط قيمة المُتَّجهين الأصليين لهذين التصنيفين (أي $(v_1+v_2)/2$). ويتم تكرار هذه الخطوة غير مرة إلى أن يتبقى تصنيف مفرد.

في الخطوة الأولى مما سبق تم التوصل إلى حساب الحِمْل المعلوماتي لكل كلمة كمية. وقد أظهرت نتيجة البحث أن أقل كلمة كمية من حيث الحِمْل المعلومات هي الكلمة الكمية "个"(١)؛ حيث كان الحمل المعلوماتي الخاص بها هو ١،٢٦٩. وهذه نتيجة غير متوقعة؛ وذلك لأن الكلمة الكمية "个" هي أكثر الكلمات الكمية التي يتفق عليها الجميع من حيث الوسطية والشيوع؛ حيث يزداد عدد الأسماء التي تتصاحب مع هذه الكلمة الكمية؛ ومن ثَمَّ فإن قدرتَها على المساهمة في عمل تصنيف للأسماء ضئيلة للغاية. وباستمرار استعراض النتيجة، نجد أن الكلمة الكمية "名" التي وصل حملها المعلوماتي إلى مقدار ٣،٣٦٣ هي كلمة كمية عامة تُستَخدَم مع الأسماء التي تُعَبِّر عن أسماء الجنس البشري. والكلمات الكمية التي تقترب قيمة الحمل المعلوماتي لها من الكلمة الكمية "名" هي "位"، و"群"، و"只"، وهذه الكلمات الكمية من الكلمات العامة التي تقترن مع عدد كبير من الأسماء. كانت الكلمات الكمية "阙"، و"题"، و"桌"، و"班" ذات أكبر قيمة للحمل المعلوماتي للسلسلة الكلامية الاسم + الكلمة الكمية؛ حيث وصلت إلى ١١،٥٢ وذلك لأن كلاً منها لم تقترن إلا باسم واحد.

(١) الكلمات الكمية في اللغة الصينية لها استخدام وظيفي للفصل بين العدد والمعدود كما أسلفنا، وليس لها معنى مقابل في اللغة العربية . (المترجم)

وعلينا أن نلاحظ أنه في الخطوة الثالثة تم استخدام قيمة المُتَّجه المُعَبِّر عن كل مجموعة من مجموعات الأسماء في التعبير عن تعريف للعدد الكلي للكلمات الكمية التي يمكن أن تتصاحب مع تلك الأسماء. وفي علم اللغة يمكن شرح تلك الطريقة على أنها: استخدام جميع الكلمات الكمية التي يمكن أن تقترن مع الأسماء في وصف الخصائص الدلالية المشتركة التي تتمتع بها تلك المجموعات من الأسماء. ولذلك فإنه نتيجة لعدم القدرة على التمييز بين مجموعات الأسماء التي تشترك في التصاحب مع كلمة كمية واحدة، يتم اتخاذ قرار منطقي بضمها معًا في تصنيف واحد، بالإضافة إلى التعبير عنها بمُتَّجه واحد. لذلك، فعلى الرغم من أن معجم الكلمات الكمية يضم عدد ١٩١٠ لاحقة اسمية وما يزيد عن ٢٠٠٠ مادة، فإنه لم يتم التقسيم إلا إلى عدد ٥٠٢ سلسلة كلامية من اسم + كلمة كمية مختلفة والمُتَجهات الخاصة بها.

إن نتيجة عملية التصنيف سالفة الذكر قد تم التوصل إليها من خلال تصور شجرة ذات فرعين. وكل طرف من أطراف تلك الشجرة يشير إلى اسم من الأسماء يشار إليه من خلال الكلمات الكمية التي تقترن معه. وأي اسمين بينهما أكبر قدر من التقارب في القيم يتم ضمهما معًا في تصنيف اسمي جديد. ويتم تكرار تلك العملية إلى أن يتم وضع جميع الأسماء في تصنيف ينتمي إلى شجرة واحدة. إن مدى مناسبة ومصداقية الفرض وطريقة الحساب التي اعتمد الكاتب عليها يتوقف على مدى قدرتنا على تقديم تصنيف دلالي مناسب انطلاقًا من تلك الشجرة التي تم التوصل إليها.

وتشير نتيجة التجربة التي قام بها خوانغ جو رن وفريقه البحثي أن الشجرة التي يقل عدد أفرعها عن أربعة أفرع، يمكن الحصول منها على

نتيجة تصنيف دلالي موثوق بها. وقد حصلوا من خلال تلك الطريقة على ٥٠-٧٥ تصنيفًا له مغزى من تصنيفات الأسماء، ويعرض الشكل ٤-٦ تصنيفين من تلك التصنيفات.

شكل (٤-٦): جزء من نتيجة التصنيف الدلالي للأسماء

a. 房子， 屋子[个，栋， 间，幢]

宿舍，校舍，房舍，精舍，官舍，公寓[栋，间，幢]

楼房，洋房[个，座，栋，间，幢]

官邸，宅邸，大厦，广厦，华厦飞别墅，古盾，大磨，庙宇，寺宇，屋宇，楼宇，宅院[座，栋，间，幢]

镖，飞镖[支，枝，枚]

冲天炮，烟斗，箭，弓箭，利箭，弩箭[支，枝]

b. 栋杆，竹竿，撑竿，钓竿，鱼竿，矢，箭矢[支，校，个，根]扫把，火把，矛，铁矛，长矛，竹蜻蜓，木棍，铁棍，警棍，烟卷[支，枝，根]

长鞭，竹鞭，教鞭，马鞭，烟，香烟，大麻烟，洋烟，长寿烟[支，枝，根，条]

栏杆，电线杆[支，枝，个，根，排]

كما أظهرت نتيجة التجربة أنه عندما تزيد أفرع الشجرة عن خمسة أفرع فإن النتيجة تشير إلى اختلافات كبيرة يتعذر معها التصنيف، وفيما يلي نقدم شرحًا لأسباب ذلك:

(١) من المحتمل ألا تكون قيمة المُتَّجه الجديد الذي يُحسَب بعد عملية الجمع بين الأسماء غير مؤكدة. وهذا الوضع يؤدي إلى اختفاء

بعض العناصر بصورة سريعة؛ وذلك لأن عدم تمييز الكلمات الكمية غير الاقترانية في أثناء حساب الحمل المعلوماتي لمجموعة الكلمات الكمية موضع الدراسة يؤدي إلى حدوث تصنيف خادع وغير واقعي للأسماء. ولذلك نجد أن هذه الطريقة لا تُمَكّننا من تمييز حالتي التضارب الدلالي وعدم الارتباط الدلالي. وإذا تحدثنا من الناحية النظرية نجد أن الفئتين الفرعيتين اللتين بينهما تضارب دلالي تُضعِف كلٌّ منهما الأخرى. وهذا يعني أنه بالنسبة إلى الفئات الجديدة التي يتم التوصل إليها فإن هذه الخصائص المحدَّدة غير ذات صلة. ومع ذلك، فإذا لم تكن هناك علاقة بين الكلمات التي لا تتصاحب مع تلك الكلمات؛ التي تم تجميعها في تصنيف واحد، فإن الخصائص الدلالية التي تم ترميزها لتوضع في فئة فرعية أخرى ما زالت هناك إمكانية لتأثرها بتلك الفئات التي تم تكوينها، ومن هنا يتضح أن قدرتها الوصفية تضعف. وحتى يمكننا حل هذه المشكلة فإن هناك حاجة إلى تصميم نموذج أكثر دقة. ولكن هذا النموذج يتطلب أن تشير المعاجم ضمن بياناتها إلى معلومات توضح الأسباب التي تمنع تصاحب كلمة كمية معينة مع مجموعة من الأسماء. ومن الناحية المنهجية فإن هذا الوضع سوف يؤدي إلى عدم إمكانية استخدام البيانات التجريبية في إثبات الفروض موضع الجدل. ولذلك فقد كان من غير المجدي التعمق في هذا الاتجاه.

(٢) دائمًا ما يكون هناك لبسٌ في عملية مراقبة الكلمات الكمية. على سبيل المثال الكلمة الكمية "条" قد تشير إلى سبع خصائص دلالية كما يلي:

(أ) أشياء طويلة ورفيعة.

(ب) حيوانات طويلة.

(ج) أشياء طويلة فوق الأرض (نفق- قناة، وما إلى ذلك).

(د) خط، بما في ذلك الخطوط التجريدية.

(هـ) قانون، لائحة، خَبَر.

(و) مصير.

(ز) أغنية. وقد اعْتَبَر خوانغ جُو رن وفريقُه البحثي في دراستهم كل كلمة كمية عبارة عن علامة وحيدة، ولم يتجهوا للتمييز بين خصائصها الدلالية المختلفة. وفائدة ذلك تكمن في سهولة المعالجة؛ ولكن من الممكن أن تَضَعَ الكلمات ذات الخصائص المختلفة في فئة واحدة؛ وذلك بسبب أن لها علامات شكلية متشابهة. ويستعين خوانغ جو رن وفريقُه البحثي حاليًا بالمعلومات الواردة في معجم للكلمات الكمية ومعجم آخر لتصاحبات الكلمات الكمية مع الأسماء؛ للحصول على علاقات تصاحب أكثر تفصيلاً والوقوف على حصر لجميع المعاني لكل كلمة كمية. وعلى هذا الأساس فقد زادت

مجموعات الأسماء وفقًا للحقول الدلالية إلى ما يزيد عن الضعف؛ حيث وصلت إلى ١٠٦١ مجموعة. وقد أدى ذلك إلى التمكّن من التمييز الصحيح لمجموعات الأسماء التي تم تصنيفها بصورة خاطئة؛ بسبب نقص تفاصيل استخدام الكلمات الكمية في أثناء عملية التحليل الأولى. وسوف يؤدي ذلك إلى نتيجة أفضل بالنسبة إلى عمليات التصنيف الدلالي للأسماء.

الفصل الخامس

استقراء إشكاليات التداخل المعجمي النحوي

عن طريق استعراض التراكيب النحوية المختلفة ذات الصلة بمجموعة من الكلمات المترادفة، يمكننا أن نُمَيِّز أكثر الكلمات قُرْبًا من حيث المعنى، بالإضافة إلى أنه من خلال استعراض أنواع الكلمات المختلفة التي ترتبط بالتركيب النحوي للمترادفات يمكننا أن نُمَيِّز بين التراكيب النحوية الأكثر تشابهًا. وهذا النوع من الدراسات يُطلَق عليه اسم: دراسات التداخل المعجمي النحوي (**Lexico-grammatical associations**).

وقد قام بيبر (**Biber .D**) بعمل استعراض للعلاقات النحوية المتعلقة بصفتين مترادفتين في المعنى بدرجة كبيرة وهما "**little**"، و"**small**"؛ وذلك في محاولة للتمييز بين هاتين الصفتين من خلال أنماط الاستخدام المختلفة[90]. وقد أظهَرَت الدراسة أنه على الرغم من أن معنى هاتين الكلمتين هو "صغير" بالإضافة إلى أنهما تُستخدمان دائمًا في موقع النعت أو المسند في الجملة، فهناك اختلاف واضح بينهما في درجة تَحَيُّز لكل منهما في مدى الاستخدام في هذين الموقعين النحويين، وهذا التَحَيُّز تجاه وظيفة نحوية ما دونًا عن الأخرى له علاقة وثيقة بسياق الاستخدام.

ففي اللغة الإنجليزية تقع النعوت الوصفية (**attributive adjectives**) قبل الاسم في الجملة، وتُستخدم في تقديم معلومات تخص ذلك الاسم. على سبيل المثال:

"The little girl next door pulled him through the fence".
"But I'm not a small person".

أما الصفات التي تأتي في موقع المسند فترد بعد الأسماء الموصولة (**copula**)، ووظيفتها تقديم معلومات عن المسند إليه داخل الجملة، على سبيل المثال:

"When she was little, she couldn't say Jessica".
"Did you think it would be too small?"

وتضم الذخيرة التي تم العمل عليها جزأين: الجزء الأول عبارة عن ذخيرة من نصوص حوارية يصل حجمُها إلى خمسة ملايين كلمة تم اجتزاؤها من ذخيرة بي إن سي (**BNC**)، والجزء الثاني عبارة عن ذخيرة من نصوص علمية يصل حجمُها إلى خمسة ملايين كلمة تم اجتزاؤها من ذخيرة لونجمان- لانكاستر (**Lancastter-Longman**). وقد تم عمل ترميز آلي لأنواع الكلمات داخل الذخيرة بأكملها؛ حيث أضيفت إلى كل كلمة معلومات تُبَيِّن نوعها النحوي، بما في ذلك علامات لتمييز النعوت الوصفية، والنعوت الخبرية.

وفيما يلي نورد النِّسَب المئوية الخاصة بهاتين الصفتين "**little**"، و"**small**" في حالات ورودهما كنعوت خبرية:

	في ذخيرة النصوص العلمية	في ذخيرة النصول الحوارية
little	%١>	%٢
small	**%13**	%٢٣

وتُظْهِر نتيجة الإحصاء أن غالبية حالات ظهور هاتين الصفتين في الذخيرتين تتجه إلى موقع النعت في الجملة وليس موقع المسند. وفي الوقت نفسه أظهرت نتائج الإحصاء أن كلمة "**small**" ظَهَرَت بنسب أكبر بكثير من "**little**" في موقع المسند؛ حيث كانت النسبة في النصوص الحوارية ٢٣%، وفي النصوص العلمية وصلت إلى ١٣%. في حين أن نسبة ظهور كلمة "**little**" في موقع النعت الخبري كانت قليلة بشكل ملحوظ (٢% في النصوص الحوارية، وأقل من ١% في النصوص العلمية).

بالإضافة إلى ذلك فقد قام بيبر بعمل استطلاع للعلاقات النحوية الخاصة بالفعلين المترادفين "**begin**"، و"**start**". ففي معظم الحالات كان من الممكن لهذين الفعلين أن يتبادلا المواقع؛ مثل:

"After the race started..."
"After the race began... "

ففي الحقيقة، إن هذين الفعلين متشابهان تمامًا في قدرتهما على تكوين علاقات نحوية؛ بمعنى أنهما متكافئان من حيث القدرة على الاقتران مع الكلمات الأخرى. فيمكن لكلا الفعلين أن يكون لهما صفة التعدي واللزوم في الوقت نفسه، على سبيل المثال:

(١) نمط التعدي: اقتران الفعل مباشرة بمفعول به مُكَوَّن من تعبيرة اسمية؛ مثل:

"Then they started/begun [leaving the quota system] "

(٢) نمط الزوم: لا يقترن بمفعول به مباشر؛ مثل:

"I had better issue a survival kit before we start/begin".

ففي حالة التعدي، إما أن يكون المفعول المباشــر تعبيــرة اســمية، أو جملة فعلية مصدرية؛ مثل "جملة فرعية -to"، أو جملة حالية تحتوي "جملــة فرعية -ing"؛ حيث تأخذ كل من "begin"، و"start" صيغتي التحول التاليتين:

(١) في نمط التعدي تحتل الجملة الحالية التــي تحتــوي "ing" مكــان المفعول به بعد الفعل مباشرة.

"They had started/begun[leaving]before I arrived".

(٢) في نمط التعدي تحتل الجملة الفعلية المصدرية "to" مكان المفعول به بعد الفعل مباشرة. وفيما يلي نعرض نتائج الدراسة الاســتقرائية التي قام بها العالم بيبر (Biber) على ذخيرة في مجــالين لغــويين مختلفين، الأول نصوص روائية (٢ مليون كلمة)، والثاني نصوص علمية (٢ مليون كلمة) وذلك للتعرف على حالات الارتباط النحوي لهذين الفعلين. وقد تم اجتزاء نصوص تلك الذخيرة بأكملهــا مــن ذخيرة لونجمان–لانكاستر Longman -Lancaster

جدول (٤–٧): نتيجة استقراء حالات الارتباط النحوي

لفعلي "begin"، و"start" في اللغة الإنجليزية

		نمط اللزوم	نمط التعدي			إجمالي
			+NP	+to- جملة فرعية	+ing- جملة فرعية	
Begin	نصوص روائية	٥٤(٢٢%)	٨(٣%)	١٧٨(٧٢%)	١٠(٤%)	٢٥٠(١٠٠%)
	نصوص علمية	٨٢(٤٣%)	٢٢(١٢%)	٦٦(٣٤%)	٢٢(١٢%)	١٩٢(١٠٠%)
Start	نصوص روائية	١٠١(٤٠%)	٥٥(٢٢%)	٥٠(٢٠%)	٤٤(١٨%)	٢٥٠(١٠٠%)
	نصوص علمية	٩١(٦٤%)	٢٢(١٦%)	٢١(١٥%)	٨(٦%)	١٤٢(١٠٠%)

وقد أظهرت الدراسة التي تَمَّت على تلك الذخيرة أن الفعلين قد ظهـرا في جميع حالات الاقتران التي تم رصدها. إلا أنه قد تبين مـن المعلومـات الإحصائية وجود طريقتين للاستخدام على درجة من الأهمية:

(١) حالات اللزوم للفعل "start" أكثر من مثيلاتها للفعل "begin".

(٢) يتجه الفعل "begin" إلى الاقتران مع المفعول به المُكَوِّن من الجملـة الفعلية المصدرية "الجملة الفرعية-to" أكثر من الفعل "start".

وقد أظهرت نتيجة الإحصاء أن نسبة ٤٠% من حالات ظهور الفعـل "start" في النصوص الروائية كانت في الحالة اللازمة، أما في النـصوص العلمية فقد وصلت إلى ٦٤%. وعلى العكس من ذلك، كانت حالات ظهـور الفعل "begin" في حالة اللزوم ٢٢% في النصوص الروائية في مقابل ٤٣% للنصوص العلمية.

عندما يظهر الفعل "start" في حالة اللزوم في النصوص العلمية يكون معناه التعبير عن بداية مرحلة معينة؛ مثل:

"... the soil formation process may start again in the fresh material".

"Blood loss started about the eighth day of infection..."

"Tillering starts about a week or earlier after broadling".

غالبًا ما يكون المسند في هذا النوع من الجمل الـذي تمثلـه الأمثلـة السابقة مُعَبِّرًا عن تصنيف اسمي، بالإضافة إلى أن الفعل في مثل هذه الجمل يتبعه حال. ففي ذخيرة النصوص الروائية كانت أغلبية حالات المسند إليـه تُشير إلى إنسان أو جماد، وغالبًا ما يأتي بعد الفعل في تلك الجمل حال. على سبيل المثال:

"As he started down the hill,he could see it... "
"... the train had started again... "

ومن هنا، نجد أن هناك حاجة إلى مزيد من الدراسات عن الحال الذي يظهر في الجمل التي تحتوي أحد تلك الأفعال اللازمة. كما أظهرت الإحصاءات أن الفعل "begin" في الذخيرتين غالبًا ما يأتي في حالة التعدي وذلك مقارنة بالفعل "artst". ففي النصوص الروائية، كانت نسبة ٧٢% من حالات ظهور الفعل "begin" يقترن بها مفعول به عبارة عن صيغة المصدر المُكَوَّنَة من "to + جملة صغرى"؛ أما في النصوص العلمية، فقد وصلت نسبة تلك الحالة إلى ٣٤% وعند عقد مقارنة نجد أن نسبة اقتران الفعل "start" بمفعول به مكون من "to + جملة صغرى" وصلت إلى ٢٠% (في النصوص الروائية) و١٥% (في النصوص العلمية).

ونظرًا إلى دقة التمييز الآلي للمفعول به "to + جملة صغرى "، فقد قام العالم بيبر بعمل دراسة على مستوى أعلى عن هذا النمط من الاستخدام في ذخيرة من عشرة ملايين كلمة تم استخراجها من ذخيرة لونجمان لانكاستر Longman-Lancaster. ويظهر الجدول ٤-٨ نتيجة تلك الدراسة. وقد أكدت الدراسة دقة النتائج السابقة بمعنى أن هناك ارتباطًا وثيقًا بين ظهور حالة التعدي للفعل "begin" وبين استخدام المفعول به المكون من "to + جملة صغرى "، وعلى الأخص في النصوص الروائية؛ حيث وصلت النسبة إلى ٦٠%؛ أما الفعل "start" فقد كانت حالات اقترانه مع أشكال التعدي الأخرى أكثر قوة.

جدول ٤-٨ نتيجة استقراء حالات الارتباط النحوي

للفعلين "begin"، و"start" في ذخيرة أكبر حجمًا

		جملة فرعية +to-	جميع أشكال التعدى الأخرى	إجمالي
Begin	نصوص روائية	١٨٧١(٦٠%)	١٢٦٧(٤٠%)	٣١٨٣(١٠٠%)
	نصوص علمية	٢٢١(٣٦%)	٣٨٧(٦٤%)	٦٠٨(١٠٠%)
Start	نصوص روائية	٢٩٤(١٧%)	١٤٦٢(٨٣%)	١٧٥٦(١٠٠%)
	نصوص علمية	٩٠(١٤%)	٥٤١(٨٦%)	٦٣١(١٠٠%)

وقد أوضح بيبر (Biber) من خلال دراسته الاستقرائية على تلك الذخيرة اللغوية أنه على الرغم من إمكانية القول بوجود تشابه كامل بين الفعلين "begin"، و"start" من حيث المعنى وأنماط التصاحب النحوي، فإن كلا منهما قد أظهر اختلافًا عند الاستخدام الواقعي مع اختلاف المجال اللغوي. وهذا يوضح أنه لا يمكن الاعتماد على الحس اللغوي لدى الإنسان في القطع (أو التوجيه) فيما يتعلق بأنماط الاستخدام؛ وأن الإنسان (بما في ذلك المتحدثين الأصليين للغة) لا يمكنه التنبؤ بوجود أنماط الاستخدام ذات الطبيعة المنظومية بصورة صحيحة، وأن أقصى ما يمكن التنبؤ به هو تحديد أي مجموعة من الأفعال ترتبط بأي نمط من أنماط الاقتران اللغوي. وعلى العكس من ذلك فقد أظهرت الدراسة الاستقرائية على الذخيرة أن الكلمات التي يبدو عليها من النظرة السطحية أنها مترادفة، إذا نظرنا إليها من ناحية أنماط الاقتران في أثناء الاستخدام اللغوي، فمن النادر التوصل إلى تطابق كامل بينها.

الفصل السادس

الدراسات المتعلقة بتنوع المستوى اللغوي

(register variation)

يتم تحديد المستويات اللغوية وفقًا لظروف الاستخدام؛ مثل الغرض من الكلام، والمضمون، ومكان الاستخدام، والقدرة التواصلية، والأسلوب المستخدم. فبالنسبة إلى أي متحدث باللغة هناك أهمية قصوى لعملية استيعاب (أو القدرة على استخدام) المستويات اللغوية المختلفة. ولا نكون مبالغين إذا قلنا: إنه لا يوجد شخص يمكنه الاكتفاء باستخدام مستوى لغوي واحد؛ بمعنى أن الناس في أثناء اليوم الواحد قد يستخدمون اللغة نفسها في التحدث أو الكتابة بأساليب متنوعة تنتمي إلى مستويات لغوية مختلفة. ومن هنا فإن الشخص يحتاج إلى التمتع بالقدرة على الاختيار السليم (أو التحول) بين المستويات اللغوية المختلفة. ففي جميع مراحل اكتساب اللغة التي يمر بها الإنسان، تكون هناك أهمية قصوى لعملية اكتساب الخصائص اللغوية لكل مستوى من مستويات الاستخدام اللغوي.

فسواء كان الغرض من ذلك هو فهم مراحل اكتساب اللغة في مستوياتها المختلفة، أو كان ذلك بغرض التوصل إلى وسائل فعَّالة يستخدمها

مدرسو اللغات في تعليم اللغة باستخدام أسلوب المستويات اللغوية، فإن كل هذا يتطلب في البداية التغلب على مشكلات وصف الخصائص اللغوية لمستويات الاستخدام المختلفة للغة، بغرض التوصل إلى التمييز الصحيح لتلك المستويات. وعلى الرغم من أن الباحثين قد انتبهوا مبكرًا إلى أهمية هذا النوع من الوصف اللغوي، فإن الواقع يؤكد أن هذا الهدف لا يمكن الوصول إليه إلا عند الاستعانة بمنهجيات الذخائر اللغوية في البحث اللغوي. ويرجع ذلك إلى أن التعمق في دراسة المستويات اللغوية المختلفة يحتاج إلى توافر العناصر الثلاثة التالية:

(١) الاعتماد على كمية هائلة من النصوص كمواد أساسية للدراسة.

(٢) التطرق إلى عدد كبير من الخصائص اللغوية.

(٣) عمل مقارنات كمية بين المستويات اللغوية المختلفة.

ومن الواضح أن جميع هذه العناصر تحتاج دائمًا إلى ذخائر لغوية كبيرة الحجم، بالإضافة إلى الحاجة إلى استخدام تقنيات التحليل المختلفة المتعلقة بهذه المهام. ففي البداية، يكون جمع عدد كبير من النصوص هو حجر الأساس لمثل هذه الدراسات؛ وذلك لأن النتائج البحثية التي يتم التوصل إليها مع نصوص قليلة تكون غير دقيقة بالقدر اللازم.

هذا بالإضافة إلى أن دراسة المستويات اللغوية التي تعتمد على أسلوب المقارنة لا يمكن أن تقدم وصفًا متعمقًا لتلك المستويات اعتمادًا على عدد قليل من الخصائص اللغوية. وعلى هذا الأساس أيضًا لا يمكن أن نقدم وصفًا

تلخيصيًّا دقيقًا للخصائص اللغوية المتعلقة بكل مستوى. وفي الحقيقية، لا يمكننا الاعتماد على خاصية لغوية بارزة في تمييز سمات مستوى لغوي معين إلا إذا كان الأمر متعلقًا باقتصار ظهور هذه الظاهرة في هذا المستوى دونًا عن غيره. وقد أثبت الواقع اللغوي أن العديد من المستويات اللغوية تتمتع بمجموعة من الخصائص اللغوية المشتركة فيما بينها؛ مثل معدلات ظهور الأسماء، والضمائر، والأفعال، والصفات، وما إلى ذلك. ولا يمكن التمييز بين هذه المستويات إلا من خلال مقارنة القيم التفصيلية لمعدلات ظهور هذه الخصائص في مستوى معين مقارنة بباقي المستويات. وبعبارة أخرى، فإن الخصائص اللغوية المحورية التي تُظهرها الفروق المنظومية في أثناء الاستخدام اللغوي من شأنها أن تقدم أدلة يمكن الاعتماد عليها في التمييز بين المستويات اللغوية المختلفة.

وفي النهاية، لا بُدَّ من الإقرار أن الدراسات القائمة على تحليل المستويات اللغوية تتطلب الاعتماد على أحد أساليب المنهج التقابلي؛ بمعنى أنها تحتاج إلى نقطة انطلاق (baseline) لعمل المقابلة، حتى يمكن الحكم على عدد مرات ظهور خاصية (أو مجموعة من الخصائص) في أحد المستويات اللغوية. فعلى سبيل المثال، أظهرت دراسة استقرائية قام بها العالم بيبر (Biber) أن الجملة الموصولة تظهر في النصوص الإنجليزية بمعدل ٢٥ مرة كل ألف كلمة، الأمر الذي اعتُبِرَ معه أن ذلك أعلى معدل تواتر؛ لأن متوسط ظهور هذا النوع من الجمل يتراوح بين ١-١٠ مرات كل ألف كلمة، ويتحدد عدد مرات الظهور التفصيلي وفقًا للمستوى اللغوي. وعلى

العكس من ذلك، إذا وجدنا أن الأسماء تظهر بمعدل ٢٠٠ مرة كل ألف كلمة في نوع معين من النصوص، وكان معدل ظهور الأسماء هو ٢٥ مرة لكل ألف كلمة، فإن ذلك يُعتبر حالة نادرة.

وفيما يلي نقدم تعريفًا للدراسة التي قام بها العالم بيبر (Biber) للتعرف على مقدار الاختلاف بين المستوى الشفهي والتحريري للغة[91]. ليس فقط لأن دراسة الفروق بين كل من المستوى التحريري والشفهي في اللغة من الموضوعات الساخنة دائمًا؛ التي يهتم بها العلماء في كل لغة، بل لأن تلك الدراسة ستتيح لنا فرصة التعرف على مجموعة الخصائص اللغوية التي استعان بها بيبر (Biber) للتمييز بين هذين المستويين اللغويين، بالإضافة إلى التعرف على منهجية التحليل متعدد الاتجاهات (dimensional multi-analysis) التي ابتكرها بيبر (Biber) واستخدمها لأول مرة.

في أثناء عمل دراسة تقابلية واسعة النطاق بين كل من المستوى التحريري والشفهي للغة، يكون أصعب ما في الأمر تحديد مجموعة الخصائص اللغوية التي سيتم على أساسها عقد المقارنة. فعلى سبيل المثال، عن طريق الاستقراء يمكننا التوصل إلى أن هناك فرقًا كبيرًا بين عدد مرات ظهور الجملة الموصولة في كل من النصوص العلمية والنصوص الحوارية في اللغة الإنجليزية. وعلى العكس من ذلك، فإن استخدام صيغة الفعل الماضي في هذين المجالين قليلة: ظهرت صيغة الفعل الماضي في النصوص العلمية بمعدل ٢٢ مرة كل ألف كلمة، وفي النصوص الحوارية ظهرت ٣٥ مرة. أما في النصوص الروائية فقد ظهرت صيغة الماضي بمعدل ٨٠ مرة كل ألف كلمة.

ووفقًا لهذا السبب، فإنه من غير الممكن الاعتماد على التوزيع النسبي لخاصية لغوية معينة في التمييز بين المستويات اللغوية المختلفة. ففي الحقيقة، أن الخصائص اللغوية التي يمكن التطرق إليها كثيرة جدًا، فضلاً عن صعوبة التكهن بأي منها يلعب دورًا مهمًا في عملية تحديد خصائص المستويات اللغوية موضع الدراسة. وفي الحقيقة أن الدراسات العلمية أظهرت أن الخصائص التي تُظهِر السمات اللغوية للنصوص الحوارية تشمل: الجمل المتقطعة (**fragmented**)، والتركيب المزجي (**contraction**)، وضمير المخاطب (**you**)، والجمل الاعتراضية (know **you**)، والأفعال غير المتصرفة (**be able to، need to، have to**)، بالإضافة إلى الجملة الفرعية التي تحتوي المكمل -wh، وما إلى ذلك. وعلى العكس من ذلك، نجد أن الخصائص التي تُظهر السمات اللغوية للنصوص العلمية تشمل: أنواعًا معينة من الكلمات تَظْهَر بصورة مكثفة داخل تلك النصوص؛ مثل الأسماء، والصفات النعتية، وأسماء الأشياء، وغير ذلك من الكلمات ذات الطابع الخاص، بالإضافة إلى صيغ المبني للمجهول من الأفعال والأبنية ذات الطبيعة الخاصة، مثل (**it is possible that**)، وما إلى ذلك.

وعلى الرغم من أن علماء اللغة قد اتفقوا على أن هناك عددًا من الأنماط اللغوية تظهر معًا في مستويات لغوية محددة، فإنه من الصعب التوصل إلى إحصاء كمي لعدد مرات ظهور هذه الأنماط اللغوية. وفي الحقيقة أن تمييز مجموعات الخصائص اللغوية التي تتصاحب معًا في المستويات اللغوية المختلفة لا يمكن أن يتحقق بشكل واقعي إلا من خلال استخدام

منهجيات الذخائر اللغوية. وقد أثبتت منهجية التحليل متعدد الاتجاهات التي طرحها العالم بيبر في ثمانينيات القرن العشرين قد أثبتت أن منهجية الذخائر اللغوية هي الأساس الذي يمكن من خلاله الانطلاق لحل هذه المشكلة.

إن الأساس الذي ينطلق منه التحليل متعدد الاتجاهات لوصف الاختلافات بين المستويات اللغوية المختلفة يعتمد على ذخيرة شفهية وأخرى تحريرية تضم عينات تغطي نطاقًا واسعًا لكل مستوى لغوي، ومن الضروري أن تُعَبِّر تلك الذخيرة بصورة كاملة عن الأنماط اللغوية الأساسية التي تتصاحب معًا في إحدى اللغات (مثل اللغة الإنجليزية على سبيل المثال). وقد استعان التحليل متعدد الاتجاهات الذي أُجْري على اللغة الإنجليزية في ثمانينيات القرن الماضي بذخيرة عامة، تتكون من عدد ٤٨١ نصًّا يحتوي ٩٦٠٠٠ كلمة. من بينها عدد ٣٤٠ نصًّا تم اختيارها من ذخيرة لوب؛ بحيث غطت تلك النصوص إلى حد كبير النصوص التحريرية من الذخيرة، مثل النصوص العلمية، وافتتاحيات الصحف، والروايات وما إلى ذلك؛ أما النصوص الباقية؛ التي وصل عددها ١٤١ نصًّا فقد تم اختيارها من ذخيرة لندن- لوند (London-Lund) الشفهية، وتضم الحوارات المباشرة، والمحاضرات العلنية، ونصوص الخطب الشفهية، وما إلى ذلك.

إن أول خطوة من خطوات التحليل متعدد الاتجاهات هي تحديد مجموعة الخصائص اللغوية التي سيتم دراستها. والغرض من هذه الخطوة هو تجميع الخصائص اللغوية على نطاق واسع؛ حيث يمكن من خلال تلك الخصائص اللغوية تقديم شرح يوضح الوظائف اللغوية لتلك النصوص. وقد

قدم بيبر قائمة بعدد ٦٧ خاصية لغوية من خلال التحليل متعدد الاتجاهـات للغة الإنجليزية، ويمكن ضمها في ستة عشر نموذجًا نحويًّا:

(١) علامة الزمن والحالة.

(٢) الحال المعبر عن المكان والزمان.

(٣) الضمير المعبر عن الاسم، والضمير المعبر عن الفعل **(pron-verb)**.

(٤) الاستفهام.

(٥) الصيغ الاسمية.

(٦) المبنى للمجهول.

(٧) صيغة الحالة.

(٨) خصائص الإضافة.

(٩) تعبيرات الجار والمجرور، والصفات، والظروف.

(١٠) المفردات الخاصة.

(١١) الأنواع النحوية للمفردات.

(١٢) الأفعال المتصرفة.

(١٣) الأفعال التي لها سمات خاصة.

(١٤) التراكيب المختصرة؛ التي يَنْدُر استخدامها.

(١٥) العطف.

(١٦) النفي.

والخطوة التالية التي اتبعها بيبر هي تصميم برمجية حاسوبية لتمييز كل خاصية لغوية داخل النص وإحصائها. مع مراعاة أن هناك عددًا من الخواص المركبة تحتاج إلى أسلوب تمييز يعتمد على التكامل بين الآلة والإنسان. هذا إلى جانب أن جميع الخصائص التي تتوصل الآلة إلى تمييزها يتم مراجعتها ومطابقتها من خلال التدخل البشري؛ وذلك لضمان دقة التمييز.

ويتضح مما سبق أن كمية البيانات التي يحصل عليها فريق التحليل من خلال عمليات الاستقراء التي تمت على الذخيرة الإنجليزية كبيرة بصورة تُلفت النظر. فالذخيرة مُكَوَّنة من عدد ٤٨١ نصًّا، وكل نص من شأنه أن يُخرج نتائج إحصائية عن معدلات تكرار ٦٧ خاصية لغوية. وحتى يمكن التوصل إلى مجموعة الخصائص اللغوية التي تشترك في الظهور داخل نص، اعتمد بيبر على أسلوب إحصائي أُطْلِق عليه اسم "تحليل العوامل" (analysis factor). وهو تقنية إحصائية تعتمد على العلاقات التبادلية (correlational) الغرض منها تمييز مجموعة المتغيرات المتشابهة من حيث نسب التوزيع داخل النص. أي أن أسلوب تحليل العوامل يعني التوصل إلى مجموعة الخصائص اللغوية التي تتجه لأن تتصاحب معًا داخل النص.

وقد أطلق بيبر لفظ "مؤشر" على مجموعة الخصائص التي تشترك في الظهور معًا؛ التي تعبر عن تغير مجال النص. على سبيل المثال، قد تكون مجموعة الخصائص المشتركة مُكَوَّنة من ضمير المتكلم، وضمير المخاطب، وأدوات الاستفهام التي تبدأ بـ -wh؛ وقد تكون مجموعة خصائص أخرى مكونة من الأسماء، وتعبيرات الجار والمجرور، والصفات النعتية وما إلى

ذلك. ومن خلال الخصائص اللغوية التي يتم الحصول عليها من أسلوب تحليل العوامل يمكن التعبير عن توزيع مجموعة الحقائق اللغوية المرتبطة بكل خاصية في كل نوع من أنواع النصوص. على سبيل المثال، عندما يضم نص معين كمية كبيرة من الأسماء، فمن المعروف أن هذا الوضع يصاحبه ظهور كمية كبيرة من تعبيرات الجار والمجرور والتعبيرات الوصفية، وعلى العكس من ذلك إذا كانت الأسماء في نص معين قليلة العدد، فإن ذلك سيؤدي إلى قلة ظهور تعبيرات الجار والمجرور والتعبيرات الوصفية.

ونظرًا إلى أن تلازُم ظهور الخصائص اللغوية يعكس السمات المشتركة بين النصوص التي تنتمي إلى حقل لغوي واحد، فإنه بعد اكتشاف الخصائص اللغوية التي تُحَدِّد "مؤشر" الحقل اللغوي، يمكننا تقديم شرح للوظيفة اللغوية التي يمثلها هذا المؤشر؛ وذلك من خلال عناصر ثلاثة هي المشهد الذي يعبر عنه النص، ووظيفة النص من الناحية الاجتماعية والمعرفية. على سبيل المثال، ظهور ضمير المتكلم وضمير المخاطَب وجمل الاستفهام المباشر والأمر بصورة كبيرة داخل النصوص الحوارية من شأنه أن يقدم شرحًا لخاصية التواصل التي يتسم بها هذا النوع من النصوص؛ وذلك لأن استخدام جمل الاستفهام المباشر وجمل الأمر تتطلب وجود شخص مستمع مطلوب منه التجاوب مع المتحدث؛ أما ضميري المتكلم والمخاطَب فيشير كل منهما إلى المتحدث والمستمع على التوالي. وعلى المنوال نفسه، فإن أسلوب القصر وبدايات الحديث الخاطئة (**false starts**) والكلمات الحقيقية المشيرة إلى أشياء عامة (مثل **thing**) جميعها لها علاقة بالعوامل التي تُقَيِّد الحوار وقت وقوعه.

وقد تمكن بيبر من خلال التحليل متعدد الاتجاهات الذي أجراه على ذخيرتين شفهية وتحريرية للغة الإنجليزية من تمييز خمسة مؤشرات للتحول النصي.

وفيما يلي نُورد مجموعة الخصائص اللغوية التي تتلازم مع المؤشر الأول؛ حيث إن كل مؤشر يتكون من خاصيتين لغويتين تتلازمان معًا في الظهور؛ بمعنى أن ظهور إحدى هاتين الخاصيتين بكثافة داخل النص يستدعي ظهور الخاصية الأخرى بوضوح، والعكس صحيح. وهاتان الخاصيتان المتكاملتان يمكن الإشارة إليهما بـ "الأصل" و"الفرع". ففي المؤشر الأول نجد أن مجموعة الخصائص اللغوية الأصل هي: وجود أفعال الشعور مثل "think"، و"feel"، وحُذِفَت كلمة "that" من الجمل الفرعية، والقصر، وصيغة الفعل المضارع، وضمير المخاطب. أما مجموعة الخصائص الفرعية فتشمل: وجود الأسماء، والكلمات الطويلة، وتعبيرات الجار والمجرور، وصيغة المقارنة، والصفات النعتية.

المؤشر الأول:			
أفعال الشعور	٠.٩٦	الأسماء	٠.٨٠-
حذف- that	٠.٩١	الكلمات الطويلة	٠.٥٨-
القصر	٠.٩٠	تعبيرات الجار والمجرور	٠.٥٤-
الفعل في صيغة المضارع	٠.٨٦	صيغة المقارنة	٠.٥٤-
ضمير المخاطب	٠.٨٦	الوصف النعتي	٠.٤٧-
...			
الأفعال المعبرة عن الإمكانية	٠.٥٠		
العطف	٠.٤٨		
الجملة الفرعية -wh	٠.٤٧		
حروف الجر التي تقع في آخر الجملة	٠.٤٣		

وأمام كل خاصية نلاحظ وجود رقم، هـذا الـرقم يعبـر عـن وزن (loading) هذه الخاصية بالنسبة إلى المؤشر الأول، وهو رقم يشير إلى قوة الارتباط (strength of the relationship) بين المؤشر الأول ووجود هـذه الخاصية. ومن هذا، يمكننا أن نلاحظ درجة تمثيل (representative) هـذه الخاصية اللغوية في هذا المؤشر. ويتراوح وزن الخاصية من ١ إلـى -١. فكلما كبرت القيمة المطلقة للوزن ارتفعت قيمة تمثيله داخل المؤشر، وعندما تصل قيمة الوزن إلى رقم واحد صحيح فإن هذا يعني الارتباط الكامل. ويبدو من الجدول أن أول خاصية في المؤشر الأول هي ارتبـاط ظهـور أفعـال الشعور مع الأسماء (حيث كـان وزن كـل منهمـا ٠,٩٦ و-٠,٨٠ علـى التوالي)، وهذا يمثل أقوى ارتباط داخل نطاق هذا المؤشر. وبالمقارنة نجـد أن الأفعال المعبرة عن الإمكانية (٠,٥٠) أكثر ارتباطًا مع الصفات النعتيـة (-٠,٤٧).

فكل خاصية لغوية لها وزنها الممثِّل لها داخل المؤشر. ونظرًا إلى أن الخصائص التي تتمتع بأوزان عالية هي تلك الخصائص التـي تعبـر عـن المؤشر بصورة أكبر من غيرها من الخصائص، لذلك فإن هذا الأمر يُعَـوَّل عليه كثيرًا في وضع تفسير لوظائف المستوى اللغوي الذي يعبر عنه ذلـك المؤشر. فغالبًا عندما تزيد القيمة المطلقة لخاصية من الخواص عـن ٠,٣٠ فإن هذه الخاصية يمكنها أن ينظر إليها بعين الاعتبار في عملية التفسير.

وعلى أساس قيمة وزن الخاصية يمكن عمل إحصاء كمي للخصائص التي تزيد قيمتها المطلقة عن ٠,٣٠ حتى يمكن حساب القيمة التي يحـصل

عليها هذا النص في أحد المؤشرات، وهذا ما يطلق عليه اختصارًا درجة المؤشر (dimention score). وعلى مستوى أعلى، يمكن عمل إحصاء كمي لإجمالي النصوص التي يتكون منها مستوى معين داخل الذخيرة بأكملها، ومن هذا الإحصاء يمكننا التوصل إلى المتوسط الحسابي لقيم المؤشر في كل مستوى لغوي. وعلى هذا الأساس يمكننا وصف الخصائص اللغوية لأي حقل من حقول اللغة، الأمر الذي يمكن معه عقد مقارنة بين أي مستويين لغويين، بالإضافة إلى عمل تحليل كامل لوظائف كل مؤشر.

الباب الخامس

تطبيقات منهجية الذخائر اللغوية

في علم اللغة الحاسوبي

لم يكن على سبيل الصدفة الظهور المفاجئ والسريع لمنهجية الذخائر اللغوية ونماذج التحليل الإحصائي للظواهر اللغوية وتحولهما لتيار أساسي لا يُستَغنى عنه في مجال الدراسات المتعلقة بمعالجة اللغات الطبيعيـــة (NLP) وعلم اللغة الحاسوبي اللذين ظهرا مع تسعينيات القرن العشرين. ونحن على ثقة كاملة أن القارئ سوف يلمَس من الأمثلـــة التـــي ســـوف نـــسوقها فـــي الصفحات التالية مدى تَمَيُّز هذه المنهجية.

الفصل الأول

الدراسات المتعلقة بإزالة اللبس

في تمييز حدود الكلمات داخل النصوص الصينية

يُعتبر اللبس في وضع حدود للكلمات الصينية والكلمات التي لم يتم تسجيلها أهم إشكاليتين تؤثران في مستوى دقة نتائج برمجيات التمييز الآلي لحدود الكلمات الصينية. وبصورة عامة، يمكننا تقسيم إشكاليات اللبس في تمييز حدود الكلمات الصينية إلى إشكاليتين فرعيتين: الأولى هي مشكلة اللبس بسبب التداخل في عملية تقسيم سلاسل الرموز؛ والثانية مشكلة اللبس بسبب التعدد الدلالي لبعض الرموز الصينية. وعند التطبيق الواقعي لعملية تمييز حدود الكلمات الصينية نجد أن غالبية حالات اللبس ترجع إلى اللبس الناتج عن التداخل في عملية تقسيم سلاسل الرموز. طبقًا لتقرير التقييم الذي تم على برمجية التمييز الآلي الجزئي لحدود الكلمات الصينية الذي تم تصميمه عام ١٩٩٥ في إطار المشروع الصيني القومي رقم ٨٦٣ لإنشاء الحواسيب الذكية[٩٢]، فإن دقة تمييز الكلمات الصينية باتباع منهجية إزالة اللبس الناتج عن التداخل في التمييز قد وصلت إلى ٧٨%، أما درجة الدقة في تمييز اللبس الناتج عن التوليد فقد وصلت إلى ٥٩%، وتشير تلك الأرقام إلى أن الدراسات المتعلقة بمشكلة تمييز اللبس في أثناء عمليات تمييز حدود الكلمات الصينية ستظل المحور الرئيس في حقل معالجة المعلومات باللغة الصينية.

وفيما يلي نعرض الدراسات المتعلقة بتمييز حدود الكلمات من خلال إزالة اللبس الناتج عن التداخل في التمييز، التي تمت في كل من جامعة شان شي وجامعة تشين خوا على ذخيرة واسعة النطاق. وحتى يمكننا التعريف بتلك الدراسات، فمن الضروري البدء بشرح بعض المصطلحات الأساسية المتعلقة بهذا المجال في اللغة الصينية، ولمزيد من المعلومات عن تعريف تلك المصطلحات، انظر المرجع رقم[٩٣].

سلسلة الكلمات التي تحتاج إلى فصل تداخل الحدود بين عناصرها يُطلَق عليها اختصارًا اسم (سلسلة متداخلة): على فرض أن لدينا سلسلة كلامية مُكوَّنَة من الكلمات A، وB، وC، وأن كلاًّ من تلك الكلمات يتكون من رمز أو عدة رموز. فإذا كانت الكلمات A، وAB، وBC، وC جميعها من الكلمات الموجودة في قائمة الكلمات الصينية المتعارف عليها، فإن السلسلة الكلامية ABC يُطلَق عليها اسم سلسلة كلامية تحتاج إلى إزالة اللبس من خلال فصل تداخل الحدود. وبطبيعة الأحوال فإن هذا النوع من السلاسل المتداخلة يحتوي حالات أكثر تعقيدًا من ذلك المثال، كما يلي: إذا كانت عملية تمييز حدود الكلمات تعتمد فقط على قائمة الكلمات المتعارَف عليها في اللغة الصينية ولا يتداخل معها إشكاليات أخرى كالإشكاليات النحوية والدلالية، فإن نتيجة التمييز ستكون مناسبة سواء كان تقطيع الرسالة الكلامية إلى C/AB أو كان BC/A. على سبيل المثال سلسلة الرموز "应用于" يمكن تقطيعها إلى "应用/于" "يُستخدم في..."، وكذلك يمكن التقطيع إلى "应/用于" "ينبغي أن يُستَخدم في..."؛ والسلسلة "可以为" يمكن تقطيعها إلى "可以/为" "يمكن من

أجل..."، وإلى "可/以为" "إلا أنه يعتقد أن...". ولذلك نُطلق على "应用于"، أو"可以为" اسم سلسلة متداخلة.

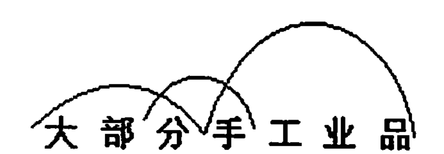
العناصر المتداخلة: الكلمات التي تتداخل في التمييز داخـل السلـسلة الكلامية يُطلَق عليها اسم عناصر متداخلة. على سـبيل المثـال، الكلمتـان "应用"، و"用于" في السلسلة "应用于" يطلق عليهما عنصران متداخلان.

طول سلسلة التداخل: عدد العناصر المتداخلة داخل السلسلة الكلاميـة يُطلَق عليه طول سلسلة التداخل. وعلى ذلك، فإن طول سلسلة التداخل فـي كل من السلسلتين الكلاميتين "应用于"، و"可以为" هو اثنان. ويتضح مـن ذلك أن طول سلسلة التداخل ينبغي أن يحتوي عنصرين متداخلين على أقـل تقدير. وبعبارة أخرى، فإن أقل قدر يمكن أن تحتويه أيـة سلـسلة كلاميـة متداخلة هو عنصران فقط. وبالمنطق نفسه، فإن أية سلسلة متداخلة لا يمكن أن يقل العدد الكلي للرموز الصينية المُكَوِّنَة لها عن ثلاثة عناصر.

محور التداخل: الجزء الذي يتكرر في الكلمتين اللتين تنتُجان من عملية التمييز يُطلَق عليه اسم محود التداخل. وعدد الرموز المُكَوِّنَة لهـذا الجـزء يُطلَق عليه طول محور التداخل. فمحور التداخل في السلسلة "应用于" هـو "用" وطول هذا المحور رمز واحد.

السلسلة ذات الطول الأكبر في التداخل: على فـرض أن S سلـسلة كلامية معينة، وفي داخل هذه السلسلة يوجد سلسلة فرعية عبارة عن سلسلة متداخلة هي S_1، فإذا كانت S لا تحتوي عنصرًا يتداخل مع S_1، فـإن S_1

يُطلَق عليها اسم السلسلة ذات الطول الأكبر في التداخل. والمثال التالي عبارة عن سلسلة متداخلة أكثر تعقيدًا:



هذه السلسلة تحتوي ثلاثة عناصر متداخلة: العنصر الأول هو "大部分" "مُعْظَم"، والثاني هو "分手" "مستقل"، والثالث هو"手工业品" "منتجات يدوية الصنع". فمحور العنصرين المتداخلين الأولين هو الرمز "分"، أما محور العنصرين الآخرين فهو الرمز "手"، وطول محور التداخل هو رمز واحد، وطول سلسلة التداخل هو ثلاثة رموز صينية. والجدير بالملاحظة أن الكلمتين "手工业" "صناعة يدوية"، و"工业品" "مصنوعات" في هذه السلسلة كلمتان متداخلتان. ولكن نظرًا إلى كون هاتين الكلمتين متضمَّنتين في إطار كلمة أكبر هي كلمة "手工业品" "مصنوعات يدوية"، فإن كلا من "手工业" "صناعة يدوية"، و"工业品" "مصنوعات" لا يمكن اعتبارهما عناصر متداخلة.

والهدف من تمييز أكبر طول لسلسلة التداخل هو وضع حدود فاصلة لهذا النوع من السلاسل؛ حيث إن السلسلة ذات الطول الأكبر لا يمكن لها أن تحتوي عناصر تداخل مع أية سلسلة رموز في الفراغ النصي المحيط بها. وهذا الوضع يُمكِّننا من اجتزاء هذه السلسلة الكلامية ومعالجتها بصورة منفردة. ففي الجملة التالية على سبيل المثال:

"经济法有普遍的强大约束力"

"القوانين الاقتصادية تفرض قيودًا كبيرة على نطاق واسع"، فنجد أن "强大约" "تقريب بصورة كبيرة"، و"强大约束力" "قيود كبيرة" سلسلتان متداخلتان، إلا أن الأولى مُتضمَّنة في الثانية. ولذلك فإن السلسلة "强大约" ليست هي السلسلة ذات الطول الأكبر في هذا المثال. ولا يبقى في هذا المثال إلا السلسلة "强大约束力" وهي التي لا تدخل في إطار سلسلة أخرى داخل هذا المثال، وهي تُعتَبر بذلك السلسلة ذات الطول الأكبر في المثال.

وفي أثناء استخراج السلاسل المتداخلة أو السلاسل ذات الطول الأكبر في التداخل من الذخائر اللغوية يتم استخدام المعادلتين الإحصائيتين "نسبة الشكل" (معدل التكرار الثابت)، و"النسبة العددية" (معدل التكرار المتغير) على التوالي وتعريف هاتين المعادلتين كما يلي:

【النسبة النوعية】: هي النسبة المئوية لعدد مرات تكرار نوع معين من أنواع السلاسل ذات الطول الأكبر في التداخل بالنسبة إلى العدد الكلي لمرات تكرار جميع الأنواع، وتُحسَب بالمعادلة التالية:

النسبة النوعية (%) =	عدد مرات تكرار نوع من أنواع السلاسل ذات الطول الأكبر في التداخل ――――――――― العدد الإجمالي لأنواع السلاسل ذات الطول الأكبر في التداخل	×١٠٠%

(معادلة ٥-١)

النسبة العددية: هي النسبة المئوية لعدد مرات ظهور نوع معين من أنواع السلاسل ذات الطول الأكبر في التداخل داخل الذخيرة بالنسبة إلى العدد الكلي لمرات ظهور جميع أنواع السلاسل ذات الطول الأكبر في التداخل. وتُحسَب بالمعادلة التالية:

$$\text{النسبة العددية (\%)} = \frac{\text{عدد مرات ظهور نوع من أنواع السلاسل ذات الطول الأكبر في التداخل}}{\text{العدد الكلي لمرات ظهور جميع أنواع السلاسل ذات الطول الأكبر في التداخل}} \times ١٠٠\%$$

معادلة (٥-٢)

أولاً: دراسة جامعة شان شي

قام كل من العالِمَين جنغ جيا خنغ، وليو كاي يينغ من جامعة شان شي الصينية بعمل دراسة مسحية عن سلاسل الرموز المتداخلة في ذخيرة لغوية مُكَوَّنة من مليون وثمانمائة ألف كلمة، ولكنهما لم يُمَيِّزَا في دراستهما بين السلاسل المتداخلة والسلاسل ذات الطول الأكبر في التداخل[94]. كما أنهما قَدَّما تعريفًا مختلفًا لطول سلسلة التداخل، ولتسهيل عملية الشرح سوف نعتمد على مجموعة المصطلحات التي سبق ذكرها وتعريفها في عرض نتيجة دراستهما.

ومن خلال الدراسة المسحية التي تمت على الذخيرة اللغوية نجح جنغ جيا خنغ، وليو كاي يينغ في استخراج عدد ٩٥٠٠ حالة لسلاسل رموز

متداخلة، وقاما ببناء بنك لهذا النوع من سلاسل الرموز. وعلى أساس ما هو متعارف عليه في المعاجم الصينية أو نتائج الإحصاءات التي تمت على ذخائر لغوية من أن حوالي ٧٠% من الكلمات الصينية تتكون من رمزين، فقد تم التركيز على أن تنحصر وحدة تقطيع السلاسل الرمزية في رمزين فقط. وقد ورد في تقريرهما أنه باعتبار كل رمزين معيارًا لتمييز الكلمات، فقد تم استخراج عدد ٨٣٧٨ سلسلة متداخلة من بين الـ ٩٥٠٠ حالة لسلاسل الرموز المتداخلة، أي (٨٨,٢%). ويوضح الجدول ٥-١ نتيجة التوزيع الإحصائي داخل هذا العدد من السلاسل وفقًا لطول سلسلة التداخل.

جدول (٥-١): النتيجة الإحصائية لسلاسل الرموز المتداخلة وفقًا لطول سلسلة التداخل

طول سلسلة التداخل	عدد السلاسل المتداخلة	النسبة النوعية	عدد مرات الظهور	النسبة العددية
٢	٤٦٤٦	%٥٥,٤	١٤٥٨١	%٦٤,١
٣	٣٤٠٩	%٤٧,٧	٧٦٣٢	%٣٣,٦
٤	٢٣١	%٢,٨	٣٨١	%١,٧
٥	٩٢	%١,١	١٤٩	%٠,٦
الإجمالي	٨٣٧٨	%١٠٠	٢٢٧٤٣	%١٠٠

من خلال البيانات الإحصائية السابقة، يمكننا ملاحظة أن مجموع عدد السلاسل التي لها طول تداخل وحدتين وثلاث وحدات بلغ نسبة ٩٦,١% من العدد الإجمالي لسلاسل التداخل، وأن عدد مرات ظهور هذين النوعين يصل

إلى ٩٧,٧%. ومن هنا، إذا تم التوصل إلى حل مشكلة تمييز السلاسل التي يصل طول تداخلها إلى وحدتين أو ثلاث وحدات، فإن ذلك من شأنه أن يرفع من نسبة الدقة في عمليات تمييز حدود الكلمات الصينية بصفة عامة. كما قام كل من جنغ وليو بعمل دراسة مَسْحِيَّة عن نتائج تمييز حدود الكلمات الصينية باعتبار الأطوال المختلفة لسلاسل التداخل، وطَرَحَا استراتيجيات التمييز الآلي القائمة على هذه الأنواع. إلا أن التقرير العلمي لدراستهما ما زال منحصرًا في إزالة التداخل القائم على اعتبار كل رمزين معيارًا لتمييز الكلمات.

(١) انحصرت نتائج تمييز السلسلة ABC باعتبار طول سلسلة التداخل وحدتين في أربع حالات كما يلي:

أ- نتيجة التمييز هي **A/BC** كما في المثال "出自己" الذي يتم تمييزه إلى "出/自己" "يُظْهِر نفسه".

ب- نتيجة التمييز هي **AB/C** كما في المثال "出现在" الذي يتم تمييزه إلى "出现/在" "يظهر في".

ج- نتيجة التمييز هي ABC، كما في المثال "传染病" "أمراض مُعْدِية".

د- عدم الاستقرار على طريقة للتقطيع.

ويوضح الجدول ٥-٢ النتيجة الإحصائية التي تمت على ذخيرة للنصوص الإخبارية حجمها مليون وثمانمائة كلمة.

جدول (٥-٢): نتيجة تمييز السلاسل المتداخلة ABC داخل الذخيرة

نتيجة التمييز	عدد السلاسل المتداخلة	النسبة النوعية	عدد مرات الظهور	النسبة العددية
A/BC	٢٣٣٠	%٥٠,١	٢٦٥٨	%٤٢,٩
AB/C	١٨٢٧	%٣٩,٣	٥٥٠٥	%٣٧,٧
ABC	٣٧٤	%٨,٠	٢٦٠٤	%١٧,٩
غير مُحَدَّد	١٢٣	%٢,٦	٢١٧	%١,٥
الإجمالي	٤٦٥٤	%١٠٠	١٤٥٨٤	%١٠٠

ويتبين من نتيجة الإحصاء السابقة أن نسبة التمييز للسلاسل المتداخلة على النمطين A/BC، وAB/C تصل إلى ٨٩,٤%، وأن نسبة عدد مرات الظهور وصلت إلى ٨٠,٦%. ومن بين هذه النسبة نجد أن تمييز نمط الاقتران AB/C من خلال التصاحبات الأكبر في اتجاه القراءة (من اليسار إلى اليمين) قد حقق نتائج تمييز صحيحة، وعلى هذا فقد تركزت صعوبة التمييز عندما يكون طول سلسلة التداخل وحدتين في نمط السلاسل المتداخلة A/BC.

(٢) عند تمييز سلسلة التداخل ABCD باعتبار أن طول التداخل ثلاث وحدات، انحصرت نتائج التقطيع في النقاط التي يوضحها الجدول التالي:

جدول (٥-٣): نتيجة تقطيع سلاسل التداخل باعتبار طول التداخل ثلاث وحدات

نتيجة التقطيع	عدد السلاسل المتداخلة	النسبة النوعية	عدد مرات الظهور	النسبة العددية
A/BC/D A/B/CD	55	%١,٦	103	%١,٤
AB/CD	3331	%٩٧,٧	7489	%٩٨,٣
ABC/D AB/C/D	7	%٠,٢	8	%٠,١
ABCD	5	%٠,٢	٥	%٠,٠٦
غير مُحَدَّد	11	%٠,٣	١١	%٠,١٤
الإجمالي	3409	%١٠٠	٧٦١٦	%١٠٠

ويُظهر الجدول الإحصائي السابق أن نمط التقطيع **AB/CD** سواء على مستوى النسبة النوعية أو على مستوى النسبة العددية يحتل نسبة ٩٨% تقريبًا من بين أنماط التقطيع الأخرى، مثل سلسلة التداخل "已经过去"، كانت نتيجة التقطيع الصحيحة لها هي "过去/已经" "تم بالفعل". لذلك، فقد اتجهت منهجية تقطيع سلاسل التداخل بطريقة ثلاثية إلى أن تكون نتيجة التقطيع على النمط **AB/CD**.

(٣) عند تمييز سلسلة التداخل **ABCDE** باعتبار أن طول التداخل أربع وحدات، لوحظ أن صعوبة التقطيع تتركز في الرموز الثلاثة الأولى. على سبيل المثال: السلسلة "为人民工作" لا يمكن الاستمرار في تمييز كلماتها بصورة صحيحة إلى مجموعة الكلمات

"为人民工作" إلا بعد النجــاح فــي تمييــز السلـسلة الداخليــة "人民/为" على أنها مُكَوَّنَة من الكلمات "为/人民/工作".

(٤) عند تمييز سلسلة التداخل ABCDEF باعتبار أن طول التــداخل خمس وحدات، كانت نتيجة التمييز الصحيحة هــي "**AB/CD/EF**". مثل "中国产品质量" التي تم تمييزها إلــى مجموعــة الكلمــات "中国/产品/质量" "جودة المنتجات الصينية".

وقد اكتشف كل من جنغ، وليو من خلال العمليات الإحــصائية التــي تمت على الذخيرة أنه من خلال تحليل عدد الــــ ٤٦٤٦ سلــسلة متداخلــة باعتبار طول التداخل وحدتين، هناك ثماني حالات اختلفت نتيجة تمييزها مع اختلاف السياق؛ مثل:

السلسلة "从小学": يتم تمييزها إلــى "从/小学" كمــا فــي الــسياق "人/三/姐妹

/从/小学/到/中学。" "انتقلت الأخوات الثلاث من المدرسة الابتدائية إلى المدرسة الإعدادية"، ويتم تمييزها إلــى "从小/学" كمــا فــي الــسياق "她/从小/学/戏剧/表演" "تعلمت العروض المسرحية منذ صغرها".

السلسلة "以北约": يتم تمييزها إلى "/以/北约" كمــا فــي الــسياق "确立/以/北约/为/核心/的/军事/。力量" "إنشاء قوة عسكرية مركزها حلف الناتو"، ويــتم تمييزهــا أيــضًا إلــى "以北/约" كمــا فــي الــسياق "兴平/市/以北/约/十五/公里"، "على بُعْد خمسين كيلو مترًا من مدينــة شينغ بينغ من جهة الشمال".

أما عند تحليل عدد الـ ٣٤٠٩ سلسلة متداخلة باعتبار طول التداخل ثلاث وحدات، فلم يحدث أي اختلاف في التمييز بسبب تغير السياق. لذلك فقد نصحا بضرورة تركيز الاهتمام بسلاسل التداخل المختلفة، بالإضافة إلى التعامل بصورة منفردة مع الأمثلة الخاصة التي تحمل تطرفًا وتؤدي إلى نتائج تحليل تختلف مع تغير السياق؛ وذلك في أثناء الدراسات المتعلقة بوضع استراتيجيات تمييز السلاسل المتداخلة.

وعلى أساس الدراسة الاستقرائية التي تمت على الذخيرة اللغوية الصينية، تَمكّن فريق البحث بجامعة شان شي من وضع قواعد تمييز السلاسل المتداخلة باعتبار أن طول التداخل وحدتان، وقد استخدموا تلك القواعد في عمل اختبار مغلق لتمييز أنواع التداخل على عدد ٤٦٤٦ سلسلة، وقد وصلت دقة التمييز في هذه العملية إلى ٨٧%. ثم استعانوا بالقواعد نفسها في عمل اختبار مفتوح على ذخيرة للنصوص الصحفية مُكَوَّنَة من مليوني كلمة، فكانت نسبة الدقة ٨١%.

فما يجب الإشارة إليه هو أن نتيجة الدراسة السابقة التي أُجْرِيَت في جامعة شان شي عبارة عن إحصاء لوحدات تقطيع النصوص الصينية بطول وحدتين للتداخل، وأن الذخيرة التي تم الاستعانة بها ذخيرة لا يتعدى نطاقها مليونًا وثمانمائة كلمة. وأن الواقع الحقيقي لظاهرة السلاسل المتداخلة أكثر تعقيدًا مما توصل إليه فريق العمل.

ثانيًا: دراسة جامعة تشينغ خوا

استعان سوين ماو سونغ وطلابُه بقائمة المفـردات TH-WL التابعة لجامعة تشينغ خوا؛ التي تضم عدد ١١٢،٩٦٧ كلمة، في استخراج جميع السلاسل المتداخلة التي وصل عددها إلى ٢٣٣،٨٨٨ نوعًا؛ وذلك من ذخيرة Rcorpus للنصوص الصحفية؛ التي يصل حجمها إلى ١٠١،٥٠٦،١٥٢ كلمة[٩٥]. وقد تراكم عدد مرات ظهور هذه السلاسل في ذخيرة Rcorpus ليصل إلى ١،٧٩٣،٣١٧ مرة، بإجمالي عدد رموز يصل إلى ٦،٥٦٦،٢٤٤، ويمثل هذا العدد من الرموز نسبة ٦,٤٧% من حجم ذخيرة Rcorpus.

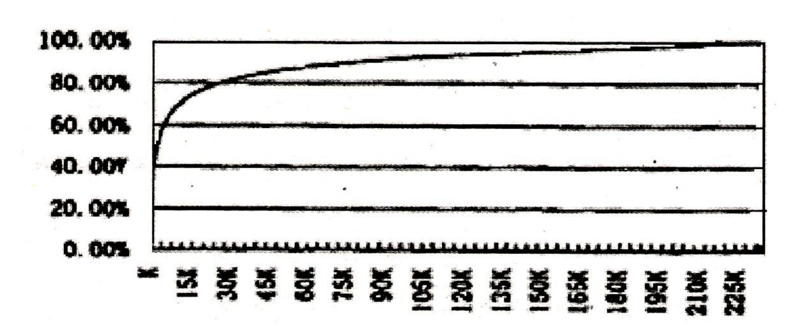
ويعرض الشكل ٥-١ المنحنى البياني $r(n)$ للسلاسل ذات الطول الأكبر في التداخل n مع مراعاة أولوية الظهور داخل الذخيرة، وبمعلومية معدل تغطية السلاسل المتداخلة r. وقد أظهرت الإحصاءات أن عدد الـ ٢٥٠٠ سلسلة المتداخلة ذات معدل التكرار الأعلى؛ التي توالى ظهورها منذ بداية الذخيرة قد تجاوزت نسبة تغطيتها ٥٠%، وأن الـ ٤٦١٩ سلسلة التي لها أكبر طول تداخل؛ التي تَوَالَى ظهورُها منذ بداية الذخيرة قد وصلت نسبة تغطيتها إلى ٥٩,٢%.

ومن أجل تجريب النتيجة التي تم التوصل إليها، قام فريق البحث بعمل مراجعة لمعدل تغطية الـ ٤٦١٩ سلسلة التي لها أكبر طول تداخل؛ التي تم التوصل إليها سابقًا على ذخيرة أطلق عليها اسم Acorpus وتضم نصوصًا صحفية، وعلمية، وعسكرية يصل حجمها إلى ستة ملايين كلمة. ويوضح

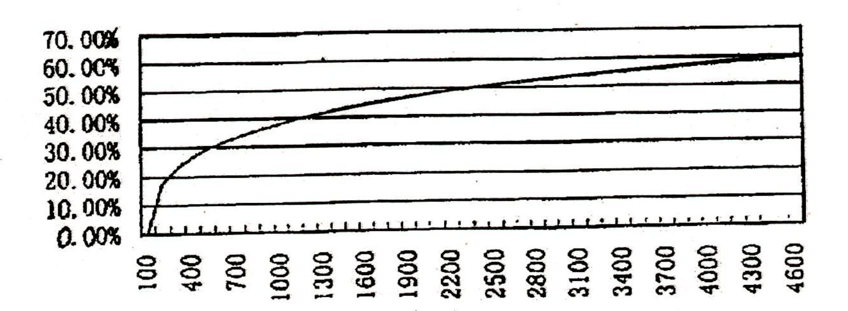
الشكل ٥-٢ أن عدد الـ ٤٦١٩ سلسلة الّتي لها أكبر طول تداخل؛ التي توالى ظهورها في ذخيرة **Rcorpus** احتفظت بنسبة التغطية نفسها تقريبًا عندما تم تطبيقها على ذخيرة **Acorpus**؛ حيث وصلت نسبة تغطيتها إلى ٥٠,٩%. وهذا الرقم، على الرغم من كونه أقل من نتيجة التغطية بذخيرة **Rcorpus**، فإن نسبة الانخفاض ما زالت محدودة بدرجة كبيرة. وهذا يعني أن نسبة التغطية التي تم الحصول عليها من ذخيرة **Rcorpus** مستقرة إلى حدٍّ كبير، وأن تأثير تَغَيُّر مجال الاستخدام اللغوي فيها ليس كبيرًا، وهذا يعني أن تلك النسبة قياسية ويمكن تعميمها.

شكل (٥-١): نسبة التغطية في ذخيرة Rcorpus

معدل تغطية السلاسل المتداخلة n مع مراعاة أولوية الظهور داخل الذخيرة بمعلومية السلاسل ذات الطول الأكبر في التداخل r



شكل (٥-٢): معدل التغطية r داخل الذخيرة Acorpus للسلاسل المتداخلة ذات الطول الأكبر n التي تتمتع بنسبة تواتر أعلى داخل ذخيرة Rcorpus



وقد استخلص سوين ماو سونغ من الـ ٤٦١٩ سلسلة التي لها أكبر طول تداخل؛ التي توالى ظهورها في ذخيرة Rcorpus ثلاث حالات للبس كما يلي:

(١) لَبْس زائف: حيث يكون هناك نتيجة واحدة لتمييز السلسلة المتداخلة.

(٢) لَبْس حقيقي من النوع ١: حيث يكون هناك نتيجتان أو أكثر لعملية التمييز.

(٣) لَبْس حقيقي من النوع ٢: جوهر السلسلة يشير إلى وجود لبس حقيقي؛ ولكن دائمًا ما يكون هناكٍ نتيجة واحدة للتمييز. وبعبارة أخرى، فإن فرص ظهور نتيجة أخرى للتمييز تكون قليلة جدًّا، لدرجة أنه يمكن إدراج هذا الصنف ومعالجته تحت تصنيف عدم وجود لبس.

ويُظْهِر الجدول ٥-٤ النتيجة الإحصائية لنسب هذه الحالات داخل الذخيرة.

جدول (٥-٤): النتيجة الإحصائية لنسب حالات اللبس

نوع اللبس	عدد السلاسل المتداخلة	النسبة النوعية
لبس زائف	٤٢٧٩	٩٢,٦%
لبس حقيقي من النوع ١	٨٥	١,٩%
لبس حقيقي من النوع ٢	٢٥٥	٥,٥%
الإجمالي	٤٦١٩	١٠٠%

والسبب في بروز عدد حالات اللبس الزائف في نتيجة الإحصاء هو أن التحليل له حالة واحدة فقط، وهذا أمر ليس له علاقة بالسياق؛ لذلك يمكن البدء بتسجيل نتائج التمييز الصحيحة لهذا النوع من الكلمات في قائمة، ولن يحتاج الأمر إلا إلى عملية معالجة بسيطة من خلال تلك القائمة حتى يمكن تمييز هذا النوع من الكلمات بصورة دقيقة داخل النصوص. ونظرًا إلى أن مجموع نسب اللبس الزائف واللبس الحقيقي من النوع ٢ قد غطت ٩٨,١% من إجمالي الـ ٤٦١٩ سلسلة التداخل ذات معدل التكرار الأعلى في ذخيرة **Rcorpus**، ونسبة ٥٨,٦% من إجمالي سلاسل الكلمات ذات الطول الأكبر في التداخل، لذلك فإن استراتيجية التمييز البسيطة التي ذكرت آنفًا تُعْتَبَر من الوسائل الفعالة بدرجة كبيرة في حل مشاكل اللبس في تمييز الكلمات داخل السلاسل المتداخلة.

وبمراعاة الخصائص الشكلية للسلاسل ذات الطول الأكبر في التداخل؛ مثل طول سلسلة التداخل، وطول الاقتران، وطول السلسلة وغيرها من العوامل، فإن ذلك من شأنه أن يؤثر تأثيرًا مباشرًا في استراتيجيات تمييز الكلمات داخل السلاسل المتداخلة. وقد قام سوين ماو سونغ في المرجع

رقم[96] بإحصاء توزيع تلك العوامل الثلاثة داخل الذخيرة اللغوية. وكانت نتيجة الإحصاء كما يلي:

(١) توزيع أطوال السلاسل ذات الطول الأكبر في التداخل داخل ذخيرة Rcorpus:

يُستَخدَم الرقمان المُسَجَّلان بين القوسين في العمود الذي يحمل عنوان أمثلة حقيقية في الجدول ٥-٥ ليشير كلٌ منهما إلى بداية العناصر المتداخلة، وطول سلسلة التداخل على التوالي. على سبيل المثال: في الصف الأول من الجدول نجد ما يلي: "项目的(0,2)(1,2)" حيث إن الرقم الأول في (0,2) يُشير إلى أول عنصر في سلسلة التداخل وهو "项目"؛ حيث يقع الرمز الأول فيه في الموضع الأول من السسلة (أي أن موقعه في السلسة = صفر) وطوله يساوي اثنين (أي أنه قد تم تمييز كلمة مُكَوَّنة من رمزين)؛ أما الرقمان (١،٢) فيشير الأول منهما إلى ثاني عنصر في سلسلة التداخل وهو "目的"؛ حيث يقع الرمز الأول فيه في الموضع الثاني من سلسلة التداخل (أي أن موقعه في السلسة=١) وطوله يساوي اثنين أيضًا (أي أنه قد تم تمييز كلمة مُكَوَّنة من رمزين).

جدول (٥-٥): توزيع أطوال السلاسل ذات الطول الأكبر في التداخل[1]

الطول الأكبر لسلسلة التداخل	عدد الأنواع المُسجَّلة	النسبة النوعية	النسبة العددية	أمثلة واقعية
٣	٧٧٨٦١	%٣٣,٢٩	%٤٩,٧٦	项目的(0,2)(1,2)，上海市(0,2)(1,2)，为人民(0,2)(1,2)，和服务(0,2)(1,2)
٤	١٠٦٢٥٦	%٤٥,٤٣	%٣٩,٨٩	在意大利(0,2)(1,3)，离退休金(3,3)(1,3)，行政区域(0,3)(2,2)
٥	٢٩٤٨١	%١٢,٦٠	%٥,٩٨	进一部到位(0,3)(2,4)，自来水龙头(0,3)(2,3)，剩余劳动力(0,4)(2,3)
٦	١٥٥٨٣	%٦,٦٦	%٣,٥٢	申请人名法院(0,3) (2,4)，自由市场经济(0,4)(2,4)
٧	٣٠٥٥	%١,٣١	%٠,٦١	少数民族自治区(0,4)(2,4)(4,3)，与此同时差不多(0,4)(3,2)(4,3)
٨	١١٩٠	%٠,٥١	%٠,١٧	主持人请不自禁地(0,3)(2,2)(3,4)(6,2)，扎扎实实地下功夫(0,4)(3,2)(4,2)(5,3)
٩	٢٨٠	%٠,١٢	%٠,٠٤	领导人民建立新中国(0,3)(2,2)(3,2)(4,2)(5,2)(6,3)

---

(١) النماذج اللغوية في هذا الجدول والجداول الشبيهة في هذا الباب للتمثيل على ظاهرة تنفرد بها اللغة الصينية تتمثل في عدم وجود فاصل بين حدود الكلمات في النصوص التحريرية، وتحديد هذا الفاصل يقوم به العقل البشري بطريقة منطقية تعتمد على السياق وتصاحُب الرموز الصينية. أما بالنسبة إلى الآلة فالأمر يحتاج إلى طُرُق استدلالية تَستَمد مقوماتها من الأداء البشري، إلى جانب الدراسات الإحصائية لمعدلات تصاحب الرموز ونسب تكرارها. إلا أن الأمر لم يستقر عند هذا الحد؛ حيث ظهرت حالات اللبس في وضع حدود للكلمات آليًا، الأمر الذي استدعى وجود مثل تلك الدراسات التي ينفرد بها هذا الفصل، التي حققت نجاحًا كبيرًا على المستوى التطبيقي. وقد اعتَبَرَ علماء اللغويات الحاسوبية باللغة الصينية هذه القضية بمثابة عنق الزجاجة في سبيل خروج اللغة الصينية التحريرية إلى مستويات المعالجة الآلية، ونود الإشارة إلى أن ما توصل إليه الصينيون من حلول، من الممكن أن يُستَرشَد به في وضع حلول لمشكلات اللغة العربية في التشكيل الآلي للكلمات الذي يُعْتَبر أيضًا عنق الزجاجة أمام اللغة العربية في مجال المعالجة الآلية. (المترجم)

青年突击队长生龙活虎 (0,5)(4,2)(5,2)(6,4)	%٠,٠٢	%٠,٠٥	١٠٨	١٠
全民所有制表演艺术团体 (0,5)(4,2)(5,2)(6,2)(7,3)(9,2)	%٠,٠١	%٠,٠٢	٥٥	١١
合法政党参与国家政治生活 (0,2)(1,2)(2,2)(3,2)(4,2)(5,2)(6,2)(7,2)(8,2)(9,2)(10,2)	%٠,٠٠	%٠,٠١	١٤	١٢
乌兹别克斯坦共和国外交部 (0,9)(8,2)(9,2)(11,2)	%٠,٠٠	%٠,٠٠	٣	١٣
提高人民生活水平息息相关 (0,2)(1,2)(2,4)(5,2)(6,2)(7,2)(8,2)(9,2)(10,4)	%٠,٠٠	%٠,٠٠	٢	١٤
	%١٠٠	%١٠٠	٢٣٣٨٨٨	الإجمالي

وتشير نتيجة الإحصاء إلى أن النسبة النوعية والنسبة العددية للسلاسل ذات الطول ٣، ٤ رموز قد وصلت إلى ٧٨,٨٢%، و٨٩,٦٥% على التوالي، مما يُثْبِتُ بما لا يدع مجالا للشك أن هذين الطولين يمثلان الطول الأكبر لسلاسل الكلمات المتداخلة داخل ذخيرة **Rcorpus**. وإذا قمنا بجمع النسب الخاصة بالسلاسل ذات الطول ٣،٤،٥،٦ معًا سنجد أن النسبة النوعية والعددية لهذه السلاسل مجتمعة هي ٩٧,٩٨% و٩٩,١٥% على التوالي. ومن الواضح أن هذه الأطوال الأربعة للسلاسل الكلامية ينبغي أن تصبح محل اهتمام العلماء والباحثين.

(٢) توزيع طول محور التداخل في ذخيرة Rcorpus:

أظهرت نتيجة الإحصاء أن العناصر المتداخلة؛ التي تتجاور حدودها في سلاسل التداخل لا يتعدى طول تداخلها أكثر من رمز واحد (وصلت النسبة النوعية لهذا الطول ٩٩,٥٧% والنسبة العددية له ٩٩,٠٤%)، وكان

من النادر ظهور عناصر متداخلة يصل طولها إلى ٢ و٣ رموز. بالإضـــافة إلى عدم وجود عناصر متداخلة يصل طولها إلى ٤ رموز على الإطلاق.

(٣) توزيع طول السلاسل ذات الطول الأكبر في الذخيرة:

جدول (٥-٦): توزيع أطوال السلاسل ذات الطول الأكبر في الذخيرة

طول التداخل	العدد النوعى لمحور التداخل	النسبة النوعية لمحور التداخل	العدد الكمى لمحور التداخل	النسبة المئوية لطول التداخل	أمثلة واقعية
١	٤٠٤١٦١	%٩٩,٥٧	٢٦٠٧٥٨٢	%٩٩,٠٤	比如如何(0,2)(1,2)(2,2)，弄虚作假发(0,4)(3,2)
٢	١٧٣٢	%٠,٤٣	٢٥٢١٢	%٠,٩٦	民族资本家(0,4)(2,3)，留洋博士生(0,2)(1,3)(2,3)
٣	١٠	%٠,٠٠	٤٢	%٠,٠٠	犹如箭在弦上(0,2)(1,4)(2,4)，现行反革命分子(0,5)(2,5)
الإجمالي	٤٠٥٩٠٣	%١٠٠,٠٠	٢٦٣٢٨٣٦	%١٠٠,٠٠	

جدول (٥-٧): توزيع طول السلاسل ذات الطول الأكبر في التداخل

طول سلسلة التداخل	عدد سلاسل التداخل	النسبة النوعية	النسبة العددية	أمثلة واقعية
٢	٩٥١٤٨	%٤٠,٦٨	%٥٧,١٨	表现在(0,2)(1,2)，留学生会(0,3)(1,3)，国民经济基础(0,4)(2,4)
٣	١١٥٤٦٧	%٤٩,٣٧	%٣٩,٩٩	任何时候(0,2)(1,2)(2,2)，革命根据地(0,2)(1,2)(1,3)
٤	١٤٦١٠	%٦,٢٥	%١,٨١	中国营养协会(0,2)(1,2)(2,3)(4,2)，我国民族资本主义(0,2)(1,2)(2,4)(4,4)
٥	٧٧٠٢	%٣,٢٩	%٠,٩٢	野外科学工作(0,2)(1,2)(2,2)(3,2)(4,2)，逐出世界杯赛(0,2)(1,2)(2,3)(4,2)(5,2)
٦	٦٤٥	%٠,٢٨	%٠,٠٦	在野生动植物资源(0,2)(1,2)(2,2)(3,3)(5,2)(6,2)
٧	٢٧٥	%٠,١٢	%٠,٠٤	进行经常性爱国主义教育(0,2)(1,2)(2,3)(4,2)(5,4)(8,2)(9,2)
٨	٢٦	%٠,٠١	%٠,٠٠	按时运抵交割地上海(0,2)(1,2)(2,2)(3,2)(4,2)(5,2)(6,2)(7,2)
٩	١١	%٠,٠٠	%٠,٠٠	城乡居民生活水平稳固(0,2)(1,2)(2,2)(3,2)(4,2)(5,2)(6,2)(7,2)(8,2)
١٠	٢	%٠,٠٠	%٠,٠٠	个国人民生活水平和美化(0,2)(1,2)(2,2)(3,2)(4,2)(5,2)(6,2)(7,2)(8,2)(9,2)
١١	٢	%٠,٠٠	%٠,٠٠	合法政党参与国家政治生活(0,2)(1,2)(2,2)(3,2)(4,2)(5,2)(6,2)(7,2)(8,2)(9,2)(10,2)
الإجمالي	٢٣٣٨٨٨	%١٠٠,٠٠	%١٠٠,٠٠	

وقد أظهرت نتيجة الإحصاء أن السلاسل ذات الطول ٢، ٣ رمزًا تحتل النسبة الأكبر، وأن النسبة النوعية لها تصل إلى ٩٠,٠٥%، والنسبة العددية لها تصل إلى ٩٧,١٧%. وأن أطول سلسلة تم مشاهدتها داخل الذخيرة التي يصل حجمها إلى مائتي مليون كلمة كانت ١١ رمزًا ولم تظهر إلا مرة واحدة فقط.

وقد أظهر مستوى أعمق من استقراء الذخيرة أن النوع الواحد من السلاسل قد يحتوي تكوينات داخلية مختلفة من حيث التركيب. ومن الواضح أن اختلاف التركيب الداخلي للسلاسل الرمزية ذات الطول الأكبر يؤدي إلى اختلاف استراتيجية التمييز. على سبيل المثال، السلسلتان "重工业区"، و"棉花生产" لهما طول واحد هو أربعة رموز صينية، إلا أن الأولى بها محورين للتداخل طول كل منهما رمزين وهما على التوالي "重工业"، و"工业区"، أما في الثانية فنجد أن بها محور تداخل يصل طوله إلى ثلاثة رموز حيث يُولِّد هذا المحور بالتالي الكلمات ثنائية الرمز "棉花"، و"花生"، و"生产" على التوالي، وعلى هذا الأساس نجد أن طول محور التداخل بهذه الكلمات يصل إلى رمز واحد.





وهناك تراكيب معينة في سلاسل التداخل من الممكن الحصول على نتيجة جيدة في تمييز كلماتها إذا تم الاعتماد على الأنواع النحوية للكلمات؛ ولكن استخدام هذه الطريقة في معالجة تراكيب أخرى لا يضمن الحصول

على نتائج مرضية. وقد قام سوين ماو سونغ بتقسيم السلاسل ذات طول التداخل الأكبر من حيث التركيب الداخلي لها إلى نوعين هما سلاسل ذات تركيب كبير، وسلاسل ذات تركيب دقيق، هذا بالإضافة إلى أنه قدم وسيلة لوصف هذين النوعين والنتائج الإحصائية الخاصة بكل منهما كما يلي:

(١) طرق التعبير عن أنواع التراكيب الكبرى للسلاسل والنتائج الإحصائية لتقسيماتها:

نظرًا إلى أن العناصر المتداخلة هي الوحدات الأساسية التي تتكون منها سلاسل التداخل، فإنه يمكن إجراء عملية تقسيم الكلمات داخل سلسلة الرموز المتداخلة S وفقًا لتلك العناصر. وبصورة أكثر تفصيلاً يعني ذلك أن كل زوج من الأرقام داخل القوسين في الجدول السابق يشير الرقم الأول منهما إلى بداية التمييز، في حين أن الرقم الثاني يشير إلى طول سلسلة التمييز داخل سلسلة التداخل S. على سبيل المثال: سلسلة التداخل "重工业区" تم تسجيل معلومات التركيب الأكبر لها على أنه (٠،٣)؛ أي أن بداية التمييز تبدأ من الرمز الذي يحمل الرتبة (٠) داخل السلسلة وهو الرمز "重" وطول الكلمة التي يتم تمييزها هي ثلاث رُتَبْ بدءًا من الرتبة (٠) وهي الرموز (重工业) وبذلك تكون الكلمة التي تم تمييزها هي (重工业) "الصناعات الثقيلة"، و(١،٣)؛ أي أن بداية التمييز تبدأ من الرمز الذي يحمل الرتبة (١) داخل السلسلة وهو الرمز "工" وطول الكلمة التي يتم تمييزها هي ثلاث رتب بدءًا من الرتبة (١) وهي الرموز (工业区) وبذلك تكون الكلمة

التي تم تمييزها هي (工业区) "مناطق الصناعة"؛ والسلسلة "棉花生产" تم تسجيل معلومات التركيب الأكبر لها على أنه (٢،٠)، (٢،١)، (٢،٢).

وعند تصنيف إجمالي السلاسل ذات التركيب الأكبر التـي احتوتهـا ذخيرة Rcorpus؛ التي وصل عددها إلى ٢٣٣،٨٨٨ سلسلة تـم الحـصول على عدد ٣١٢ تصنيفًا. وفيما يلي يعرض الجـدول ٥-٨ معلومـات عـن التركيب الأكبر لعدد ١٢ تصنيفًا منها؛ حيث يمثل هذا العدد أهم التـصنيفات التي تم التوصل إليها. وتشير الأرقام الإحصائية في هذا الجـدول إلـى أن توزيع التركيب الأكبر للسلاسل مُركّز بدرجة كبيرة؛ حيث تمثـل معلومـات التركيب (٢،٠)، و(٢،١) بالإضافة إلـى (٢،٠)، و(٢،١)، و(٢،٢) نـسبة التواجد الأكبر من بين الـ ٧٥٠ سلسلة؛ حيث يحتوي كل سجل في الجدول أحد هذه العناصر على أقل تقدير، وقد وصل إجمالي النسب النوعية والعددية لهذه التراكيب إلى ٧٣,٦٣%، و٨٤,٥٣% على التوالي.

جدول (٥–٨): توزيع التركيب الأكبر للسلاسل المتداخلة

معلومات التراكيب الكبرى داخل السلسلة	العدد النوعي	النسبة النوعية	النسبة العددية	أمثلة واقعية
(٢،٠) (٢،١)	٧٧٨٦١	%٣٣,٢٩	%٤٩,٥٧	及其他，于今天，把风车，办法则，放风筝，同行业，开发出，转化为
(٢،٠)(٢،١) (٢،٢)	٩٤٣١٥	%٤٠,٣٤	%٣٤,٧٧	中华人民，工商行政，研制成功，主要领导，产品质量，今天下午，国家规定
(٣،٠) (٢،٢)	٦٨٧٦	%٢,٩٤	%٣,٤٧	自行车厂，旅游业务，消防队员，房地产业，发电机组，合格证书，政治局面
(2,3)(2,1)(2,0)	٥٩٨١	%٢,٥٦	%١,٥٣	文化工作者，借方生产力，上报国务院，进行规范化，参加座谈会
(3,1) (2,0)	٤٨١٩	%٢,٠٦	%١,٥١	国外交部，种子公司，促进出口，上天安门，报国务院，美国务卿，和解放军
(2,1) (2,0) (2,3) (2,2)	١١٣٠٨	%٤,٨٣	%١,٣٧	国内外贸易，为主要目标，落实在行动，展现在世人，工作主要是
(3,0) (2,3)(2,2)	٦٧٨٤	%٢,٩٠	%١,٣٤	共和国土地，联合国难民，标准化工作，大部分地区，进一步调整
(٢،٠)(٢،١) (٢،٤)	٣٨٤٠	%١,٦٤	%٠,٨٨	适应市场经济，严重刑事犯罪，中国有色金属电子集团公司
(٢،٠)(٢،١) (٢،٢)(٢،٣) (٢،٤)	٦٢٢٥	%٢,٦٦	%٠,٧٥	体高产品质量，对外加工装配，集体统一经营，专职功作人员，内部分配制度
(٤،٠)(٢،٣) (٢،٤)	١٨٥٩	%٠,٧٩	%٠,٦٨	社会主要理论，古典文学名著，出租汽车行业，经济作物种植，举足轻重地位
(٤،٠)(٢،٣)	١٤٢٠	%٠,٦١	%٠,٦٠	前所未有的，高尔夫球场，乡镇企业已，别开生面的，社会保险局
(٤،٠)(٤،٢)	١٨٢	%٠,٠٨	%٠,٥٩	市场经济体制，企业集团公司，自成一家之言，星火燎原之势，风云变换莫测
أخرى	١٢٣٨٢	%٥,٣٠	%٢,٧٥	

(٢) طرق التعبير عن أنواع التراكيب الدقيقة للسلاسل والنتائج الإحصائية لتقسيماتها:

يُقصَد بالتراكيب الدقيقة داخل سلسلة الرموز كل من موضع الكلمات المحتواة داخل السلسلة وطولها (بغض النظر عن كونها تمثل عناصر متداخلة أم لا) (بما في ذلك الكلمات التي طولها رمز واحد) على سبيل المثال:

سلسلة التداخل: 重工业区

أنواع التراكيب الدقيقة التي تحتويها السلسلة: (١،٠)، (٣،٠)، (١،١)، (٢،١)، (٣،١)، (١،٢)، (٣،١)، (٢،١)، (١،٣).

سلسلة التداخل: 棉花生产

أنواع التراكيب الدقيقة التي تحتويها السلسلة: (١،٠)، (٣،٠)، (١،١)، (٢،١)، (١،٢)، (٢،٢)، (١،٣).

ومن هنا نلاحظ أن التركيب الأكبر الواحد من شأنه أن يحتوي عددًا من التراكيب الدقيقة. وبالنسبة إلى أبسط تركيبين كبيرين (٢،٠)، و(٢،١)، والذين لهما معدل التكرار الأعلى في الذخيرة نجد أنهما يحتويان الحالات الثمانية التالية من التراكيب الدقيقة:

١- (٢،٠)، (٢،١)	"浩浩森"
٢- (١،٠)(٢،٠)(٢،١)	"受贿赂"
٣- (٢،٠)(٢،١)(١،٢)	"榕树梗"
٤- (٢،٠)(٢،١)(١،٢)	"足协从"
٥- (١،٠)(٢،٠)(١،١)(٢،١)	"不过敏"
٦- (١،٠)(٢،٠)(٢،١)(١،٢)	"安徽调"
٧- (٢،٠)(١،١)(٢،١)(١،٢)	"焕发出"
٨- (١،٠)(٢،٠)(١،١)(٢،١)(١،٢)	"水平和"

وعند تصنيف إجمالي السلاسل ذات التركيب الدقيق التـــي احتوتهـــا ذخيرة Rcorpus؛ التي وصل عددها إلى ٢٣٣،٨٨٨ سلسلة تـــم الحـــصول على عدد ١،٦٦٧ تصنيف، وفيما يلي يعرض الجدول ٥-٩ معلومات عـــن أهم التصنيفات التي تم التوصل إليها؛ حيث تُمَثِّل معلومات التركيب الدقيقـــة "(١،٠)، و(٢،٠)، و(١،١)، و(٢،١)، و(١،٢)" بالإضـــــافة إلـــــى "(١،٠)، و(٢،٠)، و(١،١)، و(٢،١)، و(١،٢)، و(٢،٢)، و(١،٣)" نـــسبة التواجـــد الأكبر من بين الـــ ١٤٠ سلسلة؛ حيث يحتوي كل سجل في الجدول أحد هذه العناصر على أقل تقدير، وقد وصل إجمالي النسب النوعية والعدديـــة لهـــذه التراكيب ٧١,٣٧%، و٨٢,٣٥% علـــى التـــوالي٣٥%. وتـــشير النتـــائج الإحصائية، على الرغم من أن توزيع التراكيب الدقيقة يبدو أنه أكثر تـــشتتًا من التراكيب الكبيرة، فإننا إذا نظرنا إلى الأمر نظرة كلية ســـنجد أن هـــذا النوع من التراكيب ما زال متماسكًا بدرجة كبيرة.

جدول (٥-٩): توزيع التراكيب الدقيقة داخل سلاسل التداخل

معلومات التراكيب الدقيقة داخل السلسلة	عدد السلاسل	النسبة النوعية	النسبة العددية	أمثلة واقعية
(١،٠)(٢،٠)(٣،٠) (١،١)(١،٢)(٢،٢) (١،٣)	٤٣١٠	%١,٨٤	%٢,١٤	外交部长,交响乐团,青年人才,生活费用,所在地区,解放军队,时装表演,无线电厂,消防队员,受灾面积
(٢،١)(١،٢)(١،١) (٢،١)(١،٢)(٢،٢) (١،٣)(٢،٣)(١،٤)	١٠٩٨٨	%٤,٧٠	%١,٣٣	大会堂会见,百分之一点,充分说明了,开发生产出,地表现出来,近年来由于,生活水平和,于今年年底
(١،٠)(٢،٠)(١،١) (٢،١)(١،٢)(٢،٢) (٣،٢)(١،٣)(١،٤)	٣٥٣٥	%١,٥١	%١,٠٢	解放生产力,革命根据地,服装设计师,出生性别比,养老保险费,极端重要性,管理科学化,发展现代化
(١،٠)(٢،٠)(٣،٠) (١،١)(٢،١)(١،٢) (٢،٢)(٣،١)	١٢٣٦	%٠,٥٣	%٠,٩٤	自行车厂,流窜犯罪, 一方面对,输电线路,小学校长,地下水位,带头人和,推动力量,安全部门,面积分别
(١،٠)(٢،٠)(٣،٠) (٢،١)(١،٢)(١،٣) (٢،٣)(١،٤)	٣٦٩٢	%١,٥٨	%٠,٧٨	代表团团长,解放军官兵,成年人犯罪,平方米面积,现代化装备,地下党组织,所有制成分,共产党内部
(١،٠)(٢،٠)(١،١) (٢،١)(١،٢)(٢،٢) (١،٣)(٢،٣)(١،٤) (٢،٤)(١،٥)	٥٩٦٧	%٢,٥٥	%٠,٧٣	随着生活水平,开放大米市场,极大地方便了,没有形成规模,外汇收入超过,内部分配方式,总结交流经验,这个中心服务
(١،٠)(٢،٠)(١،١) (٢،١)(٣،١)(١،٢) (١،٣)	٢٠٤٨	%٠,٨٨	%٠,٦٢	在座谈会,有生命力,新生长点,负有心人,和解放军,对开发区,了当事人,还有赖于,本科学家,一代表团
(١،٠)(٢،٠)(١،١) (٢،١)(١،٢)(٢،٢) (٤،٢)(١،٣)(١،٤) (٢،٤)(١،٥)	١٩٩٩	%٠,٨٥	%٠,٥٧	中国运载火箭,现有生产能力,紧急电话会议,中国外汇制度,防止水土流失,生产假冒伪劣,严重水土流失,更加深入人心,
(١،٠)(٢،٠)(٤،٠)(١،١) (١،٢)(٢،٢)(١،٣)(٤،١) (٢،٤)(١،٥)	١١١	%٠,٠٥	%٠,٥٧	成人教育中心,出乎意料之外,拳头产品开发,登山运动健将,不过如此而已,增产增收节支,有限广播电台,瓢泼大雨倾盆,四面八方支援
(١،٠)(٢،٠)(٤،٠) (١،١)(١،٢)(٢،٢) (١،٣)(٢،٣)(١،٤) (٢،٤)(١،٥)	١١٣٨	%٠,٤٩	%٠,٦	技人产出水平,自力更生发展,独立自主和平,水土流失重点,大案要案情况,主观能动作用,广播电台联合

هذا وقد استخدم المرجع رقم[96] عدد ١٢٨ علامة من علامات التصنيف النحوي للكلمات لعمل تصنيف أكثر تفصيلاً للتراكيب الكبرى الدقيقة؛ على سبيل المثال: سلسلة التداخل "办法则"، و"放风筝"، و"转化为" تركيبها الأكبر هو "(٢،٠)، و(٢،١)"، إلا أن الأنواع النحوية للعناصر المتداخلة في هذه السلاسل مختلفة، وهي على التوالي كما يلي (اسم + اسم)، و(فعل + اسم)، و(فعل + فعل).

وقد تم إضافة علامات التصنيف النحوي لعناصر التداخل في عدد الـ ٢٣٣،٨٨٨ سلسلة كلامية التي تحتويها ذخيرة **Rcorpus**، ثم قُسِّمت حسب التركيب الأكبر لها فكانت النتيجة الحصول على عدد ١٦،٥٤٨ نوعًا. ويعرض الجدول (٥-١٠) معلومات عن توزيع الـ ١٢ نوعًا الأساسية من هذه التصنيفات. وعند عمل تصنيف إضافي أكثر تفصيلاً للتراكيب الكبرى في السلاسل المتداخلة حسب معلومات التصنيف النحوي للكلمات الداخلة في التركيب، وُجِد أن هناك تشتتًا في التوزيع؛ حيث يحتوي كل صنف عدد ١٤ مثالاً، وقد تجاوز إجمالي النسبة النوعية والعددية للـ ١٢ نوعًا التي وردت في البداية ٥٠%.

وبالقاعدة نفسها من الممكن إضافة علامات توضح التصنيف المحتوي لكل كلمة من كلمات سلاسل التداخل (بما في ذلك الكلمات التي تتكون من رمز واحد). على سبيل المثال، سلسلة التداخل "立法权" "السلطة التشريعية" يتم وضع علامات التركيب الدقيق ومعلومات التصنيف النحوي لعناصرها كما يلي: "(١،٠)(٢،٠)(١،١)(٢،١)+(١،٢)**(ng,ng,ng,vg,vg)**" وأصبحت

نتيجة ذلك تمييز عدد 123,356 كلمة من بين الـ 233,888 سلسلة متداخلة في الذخيرة، بمعدل لا يزيد عن رمزين لكل كلمة من الكلمات التي تم استخراجها. وقد كان عدد الكلمات التي تم الحصول عليها من خلال هذه الطريقة لتحليل التركيب الدقيق لسلاسل التداخل كبيرًا لدرجة ضعف معها المغزى الحقيقي لهذه الطريقة في التقسيم. إلا أن نتائج الإحصاءات التي تم الحصول عليها قد كشفت عن درجة كبيرة من تعقد أنماط تراكيب سلاسل التداخل، ويعتبر ذلك من العوامل التي ينبغي أخذها بعين الاعتبار في أثناء وضع التصميم التفصيلي للطرق الحاسوبية في تمييز الكلمات داخل النصوص الصينية.

جدول (٥-١٠): التراكيب الكبرى لسلاسل التداخل + توزيع التصنيفات النحوية داخل كل تركيب

التركيب الأكبر لسلسلة التداخل	التصنيف النحوي	العدد النوعي	معدل التكرار الاستاتيكي	معدل التكرار الديناميكي	أمثلة واقعية
(2,1)(2,0)	(ng,ng)	21929	%٩,٣٨	%١٢,٧٩	族人民,深层次,工人们,年历史,一时期,假发票
(2,1)(2,0)	(vg,vg)	9644	%٤,١٢	%٥,٦٨	转化为,不适应,适用于,和解决,奉献给,应承担
(2,1)(2,0) (2,2)	(ng,ng,ng)	12579	%٥,٣٨	%٥,٣٢	中华人民,产品质量,汽车工业,干部队伍,内科学会
(2,1)(2,0)	(vg,ng)	11971	%٥,١٢	%٤,٤٩	出国门,等同志,主战场,上台阶,从政治,着眼泪
(2,2)(3,0)	(ng,ng)	4602	%١,٩٧	%٢,٨٠	基督教徒,价值观念,交响乐团,摩托车队,坐标系统
(2,1)(1,0) (2,2)	(vg,ng,ng)	7578	%٣,٢٤	%٢,٥٥	执行情况,施工人员,跻身世界,实行市场,侵犯人权
(2,1)(2,0) (2,2)	(vg,vg,ng)	9123	%٣,٩٠	%٢,٥٣	与会同志,施加压力,实现形式,筹集资金,发生事故
(2,1)(2,0)	(ng,vg)	5745	%٢,٤٦	%١,٩٠	需求和,地支持,人参加,地理解,文化为,学校对
(2,1)(2,0) (2,2)	(s,ng,ng)	1761	%٠,٧٥	%١,٤٥	中国文化,长江流域,香港客人,上海市场,西藏历史
(2,1)(2,0) (2,2)	(vg,vg,vg)	4917	%٢,١٠	%١,٣٩	联合办学,公开拍卖,引起重视,建查验受，装修改造
(2,1)(2,0)	(vg,a)	970	%٠,٤١	%١,٣٩	起重要,高清楚,打破旧,和亲切,提高大,分神秘
(2,1)(2,0)	(wg,a)	965	%٠,٤١	%١,٣٥	越发展,不断档乡已经办,着实现,的确立,向来访
أخرى		142104	%٦٠,٧٦	%٥٦,٣٦	

يتضح مما سبق أن نتائج إزالة اللبس من سلاسل الكلمات المتداخلة التي تعتمد عليها أنظمة تمييز حدود الكلمات في اللغة الصينية لم تصل إلى الحد المرغوب فيه حتى الآن. والسبب الرئيس في ذلك هو أن العلماء لم يصلوا إلى سبرْ أغوار التشابك والتعقد الكامن في سلاسل الكلمات المتداخلة، بالإضافة إلى أن الطرق الحاسوبية في إزالة اللبس لم تصل إلى حد النضج بعد. إن الأعمال الاستقرائية التي قام بها فريقا البحث بكل من جامعتي شان شي وتشينغ خوا على ذخيرة كبيرة الحجم واستخدام بيانات إحصائية دقيقة كان من شأنه إزالة ساتر الغموض عن سلاسل الرموز المتداخلة، وإتاحة الفرصة لتكوين رؤية مشتركة بين الباحثين حول تلك المشكلة، بالإضافة إلى إدراك طبيعة الاختلافات التركيبية فيما بين عناصرها. وهذه النتيجة ضرورية بدرجة كبيرة في التوصل إلى طرق حاسوبية في إزالة اللبس آليًّا عن سلاسل الرموز عند تصميم تلك البرمجيات على أرض الواقع. وقد أظهرت الأعمال التي قام بها فريقا البحث أن برمجيات التمييز الآلي لحدود الكلمات القائمة على استقراءِ الذخائر المُكَوَّنة من سلاسل متداخلة كان من شأنها رفع دقة إزالة اللبس عن سلاسل الرموز المتداخلة بصورة ملموسة.

الفصل الثاني

الدراسات المتعلقة بتمييز

التعبيرات الاسمية الأساسية في اللغة الصينية

إن تمييز التعبيرات الاسمية الأساسية من الدراسات المهمة في حقل معالجة اللغات الطبيعية واسترجاع المعلومات والترجمة الآلية وغيرها من المجالات. وقد قدم العالم تشيرش (Church) تعريفًا للتعبيرات الاسمية الأساسية في اللغة الإنجليزية (baseNP) بأنها "التعبيرات التي لا تتضمن تعبيرات اسمية أخرى بداخلها"[97]. وقد اعتبر تشيرش أن تمييز التعبيرات الاسمية الأساسية بمثابة الإشكالية التي تُمكّن من عملية ترميز الحدود اليمنى واليسرى لهذه التعبيرات، ويتم تحقيق ذلك من خلال نمط العنصر N. وقد سبق أن قامت العالمة لي وين جييه من جامعة اللغة الصينية بهونج كونج باستخدام طول التعبيرة الاسمية (N-gram) في عمل تمييز للحدود الكبرى للتعبيرات الاسمية. وقد أثبتت تجربتها أن الاكتفاء باستخدام طول التعبيرة الاسمية (N-gram) الذي سبق تحديده وفقًا للتصنيف النحوي للكلمات لا يكفي لعمل تمييز صحيح للتعبيرات الاسمية داخل النصوص الصينية. هذا وقد قام العالم جاو جوين[99] من جامعة تشينغ خوا بوضع تعريف للتعبيرات الاسمية الأساسية في اللغة الصينية، بالإضافة إلى أنه اعتمد على ذخيرة استرشادية تم ترميزها يدويًا في استخدام أنماط التراكيب النحوية مع التحرر

من السياق. إلا أن الدراسة أوضحت أن ترتيب الكلمات الذي يتطابق مع أنماط التراكيب النحوية شرطًا ضروريًا لتركيب التعبيرات الاسمية الأساسية وليس شرط اكتفاء يمكن الاستغناء به عن وجود باقى الشروط. فإذا تم الاعتماد على ترتيب الكلمات داخل النصوص في التعرُف على التعبيرات الاسمية الأساسية بصورة أساسية فسوف تكون نسبة الدقة بما لا يتعدى ٤٨,٥%. لذلك إذا أردنا عمل تمييز دقيق للتعبيرات الاسمية داخل النصوص، فسوف نحتاج إلى القواعد المتعلقة بالسياق في استخراج التعبيرات الاسمية الأساسية، التي يتم التوصل إليها من خلال التدريب والتجريب المستمرين. وقد أثبتت الدراسات أن الجمع بين كل من أنماط التراكيب الأساسية للتعبيرات الاسمية baseNP وقواعد التحويل الخاصة بالسياق المحيط بالتعبيرة الاسمية baseNP من شأنه ان يُحقّق نسبة دقة في التمييز تصل إلى ٩١,١%، و٨٧,٣% في كل من الاختبارات المغلقة والحرة على التوالي.

أولاً: تعريف التعبيرات الاسمية الأساسية baseNP في اللغة الصينية

كما سبق أن أشرنا قَدَّمَ العالم تشيرش تعريفًا للتعبيرات الاسمية الأساسية على أنها "التعبيرات التي لا تتضمن تعبيرات اسمية أخرى بداخلها"، بمعنى أن تركيبها الداخلي لا يمكن أن يحتوي تعبيرات اسمية أقل. ويبدو أن هذا التعريف لا يُمكن أن يفي بأغراض معالجة اللغات الطبيعية باللغة الصينية، على سبيل المثال تعبيرات اسمية؛ مثل: "自然语言处理" "معالجة اللغات الطبيعية"، و"亚洲金融危机" "الأزمة الاقتصادية في آسيا"،

و"经济体制改革" "إصلاح النظام الاقتصادي"، وغير ذلك من التعبيـرات الصينية التي لا يمكن أن ينطبق عليها شرط عدم إمكانيـة الاحتـواء علـى تعبيرات اسمية أقل. ولكن من وجهة نظر استرجاع المعلومـات والترجمـة الآلية فإن هذه التعبيرات لها معاني خاصة، وتحتاج المعالجـة علـى أنهـا بأكملها تنتمي إلى التعبيرات الاسمية. وقد سبق أن قدم جانغ وي قوه تقسيمًا ثلاثيًا لأنواع المُحَدِّدَات التي تأتي في التعبيرات الاسمية هي: مُحَدِّدَات للتقييد، ومُحَدِّدَات للوصف، ومُحَدِّدَات للتمييز[100]. وقد قَدَّمَ جاو جوين مجموعةً من التعاريف الشكلية للتعبيرات الاسمية الأساسية (base) وفقًا لمُحَدِّدَات التقييد كما يلي:

BaseNP → baseNP + baseNP

BaseNP → baseNP + اسم | مصدر

BaseNP → مُحَدِّد للتقييد **+ baseNP**

BaseNP → مُحَدِّد للتقييد + اسم | مصدر

مُحَدِّد التقييد ← صفة | تمييز | فعل | اسم | ظرف مكان | سلسلة من الحروف اللاتينية | عدد وكلمة كمية.

ووفقًا للتعريف السابق، يمكننا تقسيم التعبيرات الاسـمية فـي اللغـة الصينية إلى نوعين أساسيين هما: تعبيرات اسمية أساسية، وتعبيرات اسـمية غير أساسية، وفيما يلي نطرح بعض الأمثلة للشرح:

جدول (٥-١١): أمثلة على التعبيرات الاسمية الأساسية والتعبيرات الاسمية غير الأساسية

تعبيرات اسمية أساسية	تعبيرات اسمية غير أساسية
复杂的特征 这台计算机 很大成就 对于形势的估计 明朝的古董 11万职工 高速发展的经济 研究与发展 老师写的评语	甲级联赛 产品结构 空中走廊 下岗女工 促销手段 太空旅行 自然语言处理 企业承包合同 第四次中东战争

ثانيًا: أنماط التراكيب النحوية في التعبيرات الاسمية الأساسية

انطلاقًا من تعريف التعبيرات الاسمية، يمكننا أن نعرف أن تلك التعبيرات ينبغي أن تلتزم في تركيبها بقواعد نحوية مُحَدَّدَة، وقد أطلق جاو جوين على تلك القواعد غير السياقية التي تُبنى على أساس من التصنيف النحوي للكلمات والعلامات الخاصة بالتعبيرات اللغوية اسم أنماط التركيب النحوي (اختصارًا الأنماط النحوية). إلا أن مزيدًا من الدراسة قد أثبت أن ترتيب الكلمات بما يتوافق مع أنماط التركيب النحوي يُعَدُّ شرطًا ضروريًا لتكوين التعبيرات الاسمية الأساسية ولكنه ليس شرطَ اكتفاء يمكن الاستغناء بوجوده عن الشروط الأخرى. فالتعبيرة اللغوية عندما يتوافق ترتيب كلماتها مع نمط التركيب النحوي قد لا تكون تعبيرة اسمية أساسية، وينقسم هذا النوع من التعبيرات إلى الحالتين التاليتين:

١- عدم وضوح الحدود: ففي داخل الجملة توجد بعض الكلمات التي يتوافق ترتيبها داخل الجملة مع أحد أنماط التركيب النحوي. وهذه

الكلمات بهذا الترتيب قد ينطبق عليها شكل نحوي، وقد لا ينطبق. انظر المثالين التاليين:

مثال ١:

技术改造是国营企业走出困境的出路。

"التحديث التكنولوجي هو الطريق لخروج الصناعة الصينية من أزمتها" [تعبيرة اسمية أساسية (baseNP)]

مثال ٢:

IBM公司宣布全面降低个人电脑的销售价格。

"أعلنت شركة آي بي إم IBM عن تخفيض شامل في أسعار الحاسبات الشخصية" [تعبيرة غير اسمية].

ففي المثال الأول، نلاحظ التعبيرة "技术/N改造/V"، وفي المثال الثاني التعبيرة "公司/N宣布/V"، كلٌ منهما يتطابق مع نمط التركيب النحوي للتعبيرات الاسمية الأساسية كما يلي: BaseNP →V+N. إلا أن الأولى تُعتَبَر تعبيرة اسمية أساسية baseNP، أما الثانية فليست تعبيرة اسمية فحسب، وإنما لا تندرج تحت أي نمط نحوي معروف.. وبعبارة أخرى فإن كلا من كلمتي "公司" "شركة"، و"宣布" "أعلن" المتجاورتين في المثال الثاني تنتميان إلى المسند إليه والمسند على التوالي، ولا يمكن أن يضمهما حدود تركيبية لتعبيرة لغوية واحدة.

٢- اختلاف نوع التعبيرة اللغوية: قد يتواجد داخل الجملة تسلسل للكلمات يتوافق في ترتيبه مع ترتيب بعض الأبنية النحوية، إلا أن هذا التسلسل قد يكون تعبيرة اسمية أساسية، أو نوعًا آخر من التعبيرات اللغوية. انظر المثالين التاليين:

مثال ٣:

今年/T大学/N毕业生/N的/U就业/V形式/N严峻/A。

"ظروف توظيف خريجي الجامعات هذا العام صعبة للغاية".

مثال ٤:

中国/N人民/N/银行/N/今天/T宣布/V降低/V利率/N。

"أعلن بنك الشعب الصيني اليوم عن تخفيض سعر الفائدة".

ترتيب الكلمات في التعبيرتين "就业/V形式/N" "ظروف توظيف"، و"降低/V利率/N" "تخفيض سعر الفائدة" يتطابق مع أحد أنماط التعبيرات الاسمية الأساسية baseNP وهو النمط BaseNP →V +N إلا أن التعبيرة الأولى تعبيرة اسمية أساسية، أما الثانية فتعبيرة فعلية.

وقد قَدَّم المرجع [٩٩] تعريفًا بالخطوتين اللتين اتبعهما العالم جاو جوين بشأن التَّعَرُّف على التعبيرات الاسمية الأساسية كما يلي:

(أ) من ذخيرة استرشادية سبق ترميز التعبيرات الاسمية الأساسية بها يدويًا ومن دون الاعتماد على السياق يتم استخراج أنماط التراكيب

النحوية للتعبيرات الاسمية، ومن النصوص موضع الدراسة التجريبية يتم ترشيح التعبيرات الاسمية baseNP التي يتطابق ترتيب الكلمات بها مع ما تم استخراجه من الذخيرة الاسترشادية؛

(ب) الاستعانة بمنهجية اكتساب طرق التحويل في استخراج التعبيرات الاسمية baseNP اعتمادًا على القواعد المتعلقة بالسياق، ومن هنا يمكن الحكم أن تراكيب الكلمات التي تم ترشيحها هي تعبيرات اسمية أساسية حقيقية.

ثالثًا: استخراج أنماط التركيب النحوي للتعبيرات الاسمية

ينقسم استخراج أنماط التركيب النحوي إلى الخطوتين التاليتين:

١- إنشاء ذخيرة لغوية تحتوي ترميزًا يدويًا للتعبيرات الاسمية الأساسية baseNP.

٢- عمل انتخاب مبدئي لأنماط التجمع بين كلمات الذخيرة، وإنشاء مجموعة من أنماط التركيب النحوي الأساسية وفقًا للمعلومات الإحصائية الموجودة بالذخيرة اللغوية موضع الدراسة.

١) ترميز التعبيرات الاسمية الأساسية baseNP داخل الذخيرة

وصل حجم الذخيرة اللغوية التي رَمَّزَها جاو جوين بطريقة يدوية إلى مائة ألف كلمة تم تمييز حدود كلماتها وإضافة رموز للتصنيف النحوي لها آليًا؛ وقد اعتمد على تعريف التعبيرات الاسمية الأساسية بالإضافة إلى

مجموعة المعايير التالية في عمل الترميز اليدوي للتعبيرات الاسمية الأساسية في تلك الذخيرة. وقد استَبْعَد من هذه القواعد لاستخراج التعبيرات الاسمية كلا من التراكيب التي تحتوي الأداة المساعدة "的"، والتعبيرات التي تحتوي حروف العطف مثل "和"، و"与"، و"及"، و"以及"، والفاصلة القصيرة "、"[1]، والكلمات المعبرة عن الزمن، والضمائر، وحروف الجر، بالإضافة إلى التراكيب المُكَوَّنة من الأعداد والكلمات الكمية، وما إلى ذلك.

٢) الأنماط النحوية الأساسية للتعبيرات الاسمية baseNP

على أساس من التصنيف النحوي للكلمات والمعلومات الخاصة بمقاطعها الصوتية قام العالم جاو جوين باستخراج عدد ٤٠٧ نمط تركيب نحوي لتعبيرة اسمية أساسية من الذخيرة التي سبق ترميزها يدويًا؛ التي اعتبرها بمثابة ذخيرة استرشادية، ومن بين هذا العدد كان هناك عدد ٦٤ نمطًا تكرر ظهوره داخل الذخيرة أكثر من خمس مرات، بما يُغَطي نسبة ٩٨,٦% من التعبيرات الاسمية الأساسية داخل الذخيرة. وقد أطْلَقَ جاو جوين على الـ ٦٤ نمطًا المشار إليها اسم أنماط التركيب النحوي الأساسية. ويُقَدِّم الجدول التالي رقم ٥-١٢ عرضًا لأهم أنماط التركيب النحوي الأساسية، وتتكون العلامات المُسَجَّلة أمام كل تركيب نحوي من جزأين: الأول مكتوب بالحروف الإنجليزية الكبيرة للتعبير عن التصنيف النحوي للكلمة، والثاني عبارة عن رقم مُلاصِق للحروف يشير إلى عدد الرموز

(1) الفاصلة القصيرة إحدى علامات الترقيم التي تنفرد بها اللغة الصينية وتُستَخدَم في الفصل بين العناصر المتساوية في الرتبة داخل الجملة الصينية. (المترجم)

المكوِّنة للكلمة. على سبيل المثال: نمط التركيب النحوي **baseNP** الذي يحتوي العلامات "NG2+VN2" يُعَبِّر عن كلمة اسمية مكوَّنة من مقطعين صوتيين يُشار إليها بالعلامة (NG2)، ومَصْدَر مُكَوَّن من مقطعين صوتيين يشار إليهما بالعلامة (VN2).

جدول (٥-١٢): الأنماط النحوية الأساسية للتعبيرات الاسمية ومكوناتها

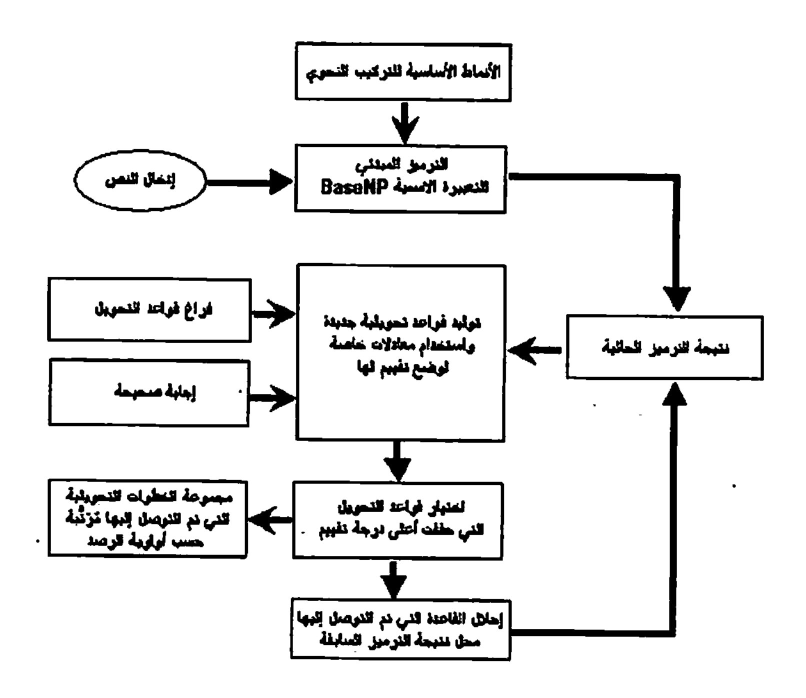
نمط التركيب النحوي	مثال توضيحي	نمط التركيب النحوي	مثال توضيحي
VN2+NG2	教育理论 调查报告	NG2+NG2+VN2	产品结构调整 住房制度改革
VGN2+NG2	委托方式 打击力度	NG2+VN2+NG2	情报检索方法 概率标引模型
VG02+NG2	下岗女工 促销手段	NG2+NG2+NG2	脂肪分子结构 热带草原气候
S2+NG2	空中走廊 海底光缆	NG2+NG2+VN2+NG2	政治体制改革促进 思想品德教育理论
XCH+NG2	0157 病毒	NG2+NG2+VG02+NG2	经济信息流通方式 航天飞机飞行计划

وقد أظهرت الإحصائيات أنه إذا تم ترميز جميع التراكيب التي تتطابق مع أنماط التركيب النحوي على أنها تعبيرات اسمية فسوف يكون معدل التغطية حوالي ٩٨,٦%، إلا أن نسبة الدقة في التمييز لن تتعدَ ٤٨,٥%. وهذا يشير إلى أن الاكتفاء بالاعتماد على الأنماط المُتَحَرِّرة من السياق لا يحل عملية الغموض في تمييز حدود التعبيرات الاسمية الأساسية وأنواع التعبيرات اللغوية.

رابعًا: قواعد تمييز التعبيرات الاسمية اعتمادًا على السياق

قدّم العالم بريل (Brill) منهجية اكتساب طرق التحويل التي يتم دعمها من خلال الخطأ[101]، وقد سبق أن قام العالم رامشاو (Ramshaw) بتطبيق هذه المنهجية في تمييز حدود التعبيرات اللغوية داخل النصوص الإنجليزية[102]. وقد استعان جاو جوين بمنهجية اكتساب طرق التحويل الخاصة ببريل في استخراج القواعد التي تُمكّن من تمييز التعبيرات الاسمية الأساسية اعتمادًا على السياق. ويوضح الشكل التالي رقم ٥-٣ طريقة الحساب التي يتم على أساسها تمييز التعبيرات الاسمية اعتمادًا على السياق:

شكل (٥-٣): رسم توضيحي يبين أنماط تمييز التعبيرات الاسمية baseNP اعتمادًا على قواعد التحويل



فالبداية تكون عن طريق عمل ترميز مبدئي للتعبيرات الاسمية الموجودة في النص الذي يتم إدخالها وفقًا لأنماط التركيب النحوي للتعبيرات الاسمية، ثم تُقارَن نتيجة الترميز المبدئي مع النتيجة الصحيحة؛ وذلك من أجل اكتشاف الأخطاء في عملية الترميز المبدئي. ومن هنا يتم الاستدلال على أنماط القواعد التحويلية المرتبطة بالسياق مقدمًا. وتُستَخدَم تلك القواعد في عمل تحويلٍ لنتائج الترميز الحالية بأكملها، بالإضافة إلى تخزين القواعد الجديدة التي تم التوصل إليها حسب أولوياتها ضمن مجموعة القواعد التي سبق رصدها. وبتكرار العمليات السابقة يتم التعلُّم من الخطأ في اكتساب القواعد التحويلية المتعلقة بالسياق واحدة تلو الأخرى. خلاصة القول: إن منهجية التعلم التحويلي هذه تتكون من ثلاث وحدات:

١) وحدة الترميز المبدئي

حيث يتم الاستعانة بالنماذج الأساسية في عمل ترميز أوّلي للتعبيرات الاسمية**baseNP** ، وفي هذه المرحلة يتم الحصول على مجموعة من التعبيرات الاسمية المُرَشَّحة. وتكون مراحل الترميز كما يلي:

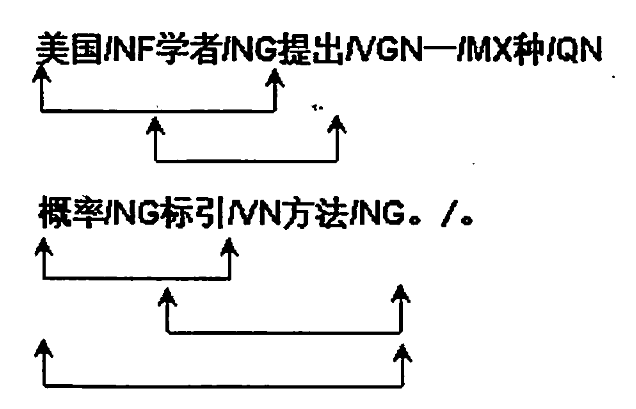
يتم تسجيل رتبة كل كلمة من محتوى النص على النمط w_i كما تُسَجَّل لكل كلمة علامة التصنيف النحوي الخاصة بها على النمط t_i، وبذلك يتحول النص المُدخل إلى سلسلة العلامات التالية:

$$w_1 / t_1 w_2 t_2 ... w_i t_i ... w_j / t_j ... w_{N-1} / t_{N-1} w_N / t_N$$

إذا كانت مثل هذه القاعدة غير السياقية موجودة ضمن الأنماط الأساسية فإنه يتم التعامل مع سلسلة الكلمات على أنها تعبيرة اسمية على النمط التالي:

$t_i...t_j, \rightarrow baseNP$

بالإضافة إلى وضع ترميز مبدئي لعناصر تلك السلسلة $w_i/t_i...w_j/t_j$ على أنها تعبيرة اسمية baseNP. على سبيل المثال:



ففي المثال السابق، نجد أن الخطوط الأفقية المُستَخدمة يشير كل منها إلى ست سلاسل لتعبيرات اسمية baseNP منتخبة، ولم يُستَقر منها بشكل نهائي إلا على التعبيرتين الموضحتين في المثال.

٢) نماذج قواعد التحويل

كل نموذج من نماذج قواعد التحويل يتكون من عنصرين رئيسين، أحدهما يختص بخطوات التحويل، والآخر يختص بتحديد نقطة البداية. فبالنسبة إلى خطوات التحويل يتم من خلالها تحديث نتيجة الترميز التي تمت على النص سابقًا. إن خطوات الترميز في هذه المرحلة تهتم بالنظر في ثلاثة

عناصر، أولها علامات الترميز التي تم وضعها للتعبيرات الاسمية المُرَشَّحة في الخطوة السابقة، أما ثانيها وثالثها فيختص بتأكيد الرموز المستخدمة أو رفضها؛ وذلك وفقًا للخطوات التفصيلية التالية:

(أ) خطوة التحويل الأولى: وَضْع علامة قبول أمام علامات الترميز المُرشحة لسلسلة الكلمات W.

(ب) خطوة التحويل الثانية: وضع علامة رفض أمام علامات الترميز المُرشحة لسلسلة الكلمات W.

(ج) خطوة التحويل الثالثة: رفض علامة القبول لسلسلة الكلمات W.

(د) خطوة التحويل الرابعة: قبول علامة الرفض لسلسلة الكلمات W.

وتتوقف شروط تنشيط قواعد التحويل على الكلمتين السابقتين للسلسلة الكلامية بالإضافة إلى الكلمة اللاحقة لها. وهذه الكلمات المجاورة للسلسلة يتم النظر إلى تصنيفها النحوي وحقلها الدلالي وعدد مقاطعها الصوتية وما إلى ذلك من الخصائص؛ وفي الوقت نفسه يتم النظر بعين الاعتبار إلى نمط التصنيف الذي تنتمي إليه سلسلة الكلمات موضع المعالجة. ولمزيد من التفصيل نورد فيما يلي عدد الـ ٢٠ شرطًا المستخدمة في تنشيط قواعد التحويل:

; $POS(p_{-1}) = t$

; $POS(p_1) = t$

; $POS(p_{-2}) = t$

; $SENSE(p_{-1}) = s$

; $SENSE(P_1)$

; $SENSE(P-2\ ) = s$

; $SYL(p_{-1}) = x$

; $POS(p_{-1}) = t_1.AND.POS(p_1) = t_2$

; $POS(p_{-2}) = t_1.AND.POS(p_{-1}) = t_2$

$POS(p_{-1}) = t_1.AND.SENSE(p_{-1}) = s_2$

; $SP(W) = m.AND.POS(p_{-1}) = t$

; $SP(W) = m.AND.POS(p_1) = t$

; $SP(W) = m.AND.POS(p_2) = t$

$POS(p_{-1}) = t_1.AND.POS(BEGIN(W)) = t_2$;

; $POS(p_1) = t_1.AND.POS(END(W)) = t_2$

; $SENSE(p_{-1}) = s_1.AND.SENSE(BEGIN(W)) = s_2$

$POS(p_{-1}) = t.AND.SENSE(p_{-2}) = s_1.AND.$

; $SENSE(BEGIN(W)) = s_2$

; $SENSE(p_{-1}) = s_1.AND.SENSE(END(W)) = s_2$

$POS(p_{-1}) = t.AND.SENSE(p_{-2}) = s_1.AND.$

; $SENSE(END(W)) = s_2$

; $SENSE(p_1) = s_1.AND.SENSE(END(W)) = s_2$

حيث تشير W فيما سبق إلى سلسلة الكلمات المُرَشَّحة؛ التـــي تخـــضع لعملية المعالجة، أما p_{-2}، p_{-1}، p_1 فتُشير كلٌ منها على التوالي إلى الكلمـــة التي تسبق السلسلة بمكانين، والكلمة التي تسبق السلسلة بمكان واحد، والكلمة التي تلي السلسلة الكلامية W؛ أما $POS(p)$، $SENSE(p)$، $SYL(p)$ فتُشير كلٌ منها إلى خصائص الكلمة التي في الموضع p مـــن حيـــث التـــصنيف النحوي، والحقل الدلالي، وعدد المقاطع الصوتية على التوالي، هذا بالإضافة إلى أنه يتم استخدام رموز الحقل الدلالي وفقًا للتصنيفات الدلاليـــة الكبـــرى والوسطى والصغرى حسب ما ورد في معجـــم ((غابـــة المترادفـــات))(١). أما $SP(W)$ فتشير إلى نمط التركيب النحوي الذي تنتمي إليه السلسة موضع المعالجة، وتشير كل من $BEGIN(W)$، $END(W)$ إلـــى الكلمـــة الأولـــى والكلمة الأخيرة في السلسلة على التوالي. وبالإضافة إلى ذلك، فإنه إذا كانت الكلمة التي في الموضع $\dot{p}$ قد انضمت لتعبيرة لغوية اسمية فيـــتم ترميزهـــا على النمط $POS(p) = BN$.

وعلينا ملاحظة أن كلاً من الحروف t، وx، و s في شروط تنـــشيط قواعد التحويل السابقة يشير إلى التصنيف النحوي للكلمة، وعـــدد المقـــاطع الصوتية، ورمز الحقل الدلالي على التوالي، ويتم وضع تلك القيم من خـــلال التدريب العملي على نصوص فعلية. لذلك، فإنه قبـــل إقـــرار تحديـــد هـــذه

---

(1) معجم (غابة المترادفات) تأليف مي جيا جو وآخرين، نُشِرَ عام ١٩٨٣ بـــدار نـــشر المؤلفات المعجمية بشنغهاي، وكان الهدف الأساسي من تأليف هذا المعجم جمع أكبـــر عدد من الكلمات المترادفة وإتاحتها للعاملين في مجـــال الترجمـــة والتـــأليف باللغـــة الصينية. (المترجم)

المتغيرات تظل قواعد التحويل قيد انتظار التقييم. وتُحدِّد مجموعة قواعد التحويل تلك مجموعة الاحتمالات الممكنة الخاصة بعمليات التحويل، وهذا ما يُطلَق عليه فراغ قواعد التحويل. والهدف من تعلم عمليات التحويل واكتسابها هو تحديد قائمة المتغيرات الخاصة بأنماط القواعد النحوية وذلك من خلال ذخيرة تم ترميزها اعتمادًا على العنصر البشري، الأمر الذي يُحَوِّل تلك القائمة إلى قواعد تحويلية يمكن تنفيذها من خلال برمجية حاسوبية. ولذلك، فمن الضروري هنا أن نُميِّز بين مفهومين مختلفين، الأول يخص نماذج القواعد المستخدمة في التحويل، والثاني هو قواعد التحويل نفسها.

٣) مُتوالِيات[١] التقييم

حتى يمكن اختيار قواعد التحويل الأفضل بالنسبة إلى نتيجة ترميز الذخيرة التجريبية، تكون هناك حاجة إلى وضع مُتوالِية للتقييم يُمكِنُها وضع درجات تقييمية لقواعد التحويل التي يتم التوصل إليها. فعلى فرض استخدام قاعدة تحويل ما يُرْمَز لها بالرمز r في النص الذي يخضع للمعالجة حاليًا، فإذا كانت نسبة دقة التمييز التي تم التوصل إليها من تطبيق هذه القاعدة على نص آخر أعلى ما يمكن، فإن هذه القاعدة تأخذ درجة تقييم مرتفعة، بالإضافة

---

(1) استُخدِمَت كلمة "متوالية" هنا للإشارة إلى ما يقابل كلمة function في اللغة الإنجليزية، وتَعني في علوم الحاسب حزْمَة من الأوامر تُكتَب بلغة برمجة وتُدمج مع برنامج متكامل للقيام بمهمة معينة، وعند استدعائها يتم تنفيذ محتواها من أسطر البرمجة بشكل تراتبي لتنفيذ المهمة التي صُمِّمَت من أجلها، وقد ارتضيت هذه الترجمة بدلا من كلمة "وظيفة"، أو "روتين فرعي" المستخدمتين في هذا المجال، أو "دالة" المُستَخدَمة في مجال الهندسة التحليلية. (المترجم)

إلى ترقيتها لتنضم إلى القواعد المستخدمة في التحويل حاليًا. وبعبارة أخرى، إذا أدت قاعدة التحويل r إلى تعديل علامات الرفض الموجودة في النص بعدد مرات قبول يصل إلى القيمة $C(r)$، وفي الوقت نفسه أدَّت إلى تعديل علامات القبول الموجودة في النص بعدد مرات رفض تصل إلى القيمة $E(r)$، فإن متوالية التقييم ستُعطِي درجة تقييم لتلك القاعدة من خلال المعادلة التالية:

$$F(r) = C(r) - E(r)$$

معادلة (٥-١)

خامسًا: طرق حساب القواعد الخاصة بتعلُّم الترميز

الغرض من التعرف على العمليات الحاسوبية الخاصة بتعلُّم الترميز هو محاولة عمل توليد آلي لمجموعة القواعد السياقية التي يتم تنفيذها بالترتيب وذلك عن طريق الاستعانة بفراغ قواعد التحويل التي سبق التوصل إليها. ففي كل خطوة من خطوات العملية الحسابية، تقوم آلية التعلم باستعراض جميع نماذج التحويل التي تنطبق عليها شروط التنشيط، وتقوم بمسح نتيجة الترميز التي تم التوصل إليها واستبدالها بمحصلة الترميز الأخيرة، مع الاستعانة بمتوالية التقييم لوضع درجة لها، واعتبار القاعدة التحويلية التي تُحرِز أعلى النتائج هي القاعدة التي يتم التوصل إليها من خلال تلك الدورة، هذا بالإضافة إلى فَهْرَستِها حسب أولوية التوصل إليها ضمن مجموعة القواعد المُحَقَّقة (القاعدة التي يتم التوصل إليها من خلال الدورة الأولي توضع في مقدمة المجموعة، والقاعدة التي يتم التوصل إليها

من خلال الدورة الثانية توضع في المركز الثاني وهكذا). وبعد ذلك، تَحِل هذه القاعدة الجديدة محل علامات الترميز الموجودة في النص موضع المعالجة، وهكذا يتم الحصول على ملف جديد، ثم الدخول في دورة جديدة. وتستمر عملية التعلُّم على هذا المنوال، إلى أن تعجز الدرجات التي تُحَقِّقُها القواعد التحويلية عن الوصول إلى قيمة معينة يتم تحديدها سابقًا. وكما ذُكر آنفًا تُفَهْرَس القواعد التحويلية التي يتم التوصل إليها من خلال عملية التعلم حسب أولوية التوصل إليها؛ حيث تتصدر القواعد التي يتم التوصل إليها مقدمًا القواعد التي تتحقق في المراحل التالية. فعند عمل ترميز للتعبيرات الاسمية داخل نص من النصوص، ينبغي البدء باستخدام أنماط ترميز مبدئية لوضع رموز للنص بشكل مبدئي، ثم يتم استبدال تلك الرموز، واحدًا تلو الآخر، بالقاعدة النهائية من مجموعة القواعد التحويلية.

ويمكننا تلخيص طرق حساب عملية تعلم القواعد التحويلية كما يلي:

على فرض أن C ذخيرة لغوية لم يتم عمل ترميز للتعبيرات الاسمية baseNP بها، وأن C_A ذخيرة تم ترميز التعبيرات الاسمية بها، وأن TS هي مجموعة القواعد التحويلية مُرَتَّبة، وأن مجموعة القواعد التحويلية تكون فارغة في البداية أي أن:

$$TS = 0$$

أولاً: يتم تطبيق نماذج الأبنية النَّحْوِية الأساسية للتعبيرات الاسمية في عمل ترميز أوَّلي للذخيرة C؛ ومن ثَمَّ الحصول على نص مُرَمَّز يُشار إليه بالرمز $C^{[0]}$.

ثانيًا: تكرار الخطوات التالية إلى أن يُصبِح من غير الممكن العثور على قواعد تحويلية r، وبحيث يمكن أن تكون $F(r) > T$ (حيث تشير T إلى قيمة مُحددة سابقًا).

الدورة رقم i (حيث $i = 0,1,2,...$)

١) مقارنة $C^{[i]}$ مع C_A، والبحث عن حالات الترميز الخاطئة $E^{[i]}$ في $C^{[i]}$.

٢) التحرك انطلاقًا من $E^{[i]}$ في البحث داخل فراغ قواعد التحويل عن أفضل قاعدة للتحويل $r^{[i]}$، وإعطاء $r^{[i]}$ أعلى درجة تقييم كما يلي:

$r^{[i]} = \arg\max F(r)$;

٣) إضافة $r^{[i]}$ إلى ذيل القائمة TS، بالإضافة إلى إحلال $r^{[i]}$ مكان $C^{[i-1]}$ والحصول بذلك على $C^{[i]}$.

سادسًا: نتيجة التجربة

انقسمت التجربة إلى الأجزاء الثلاثة التالية:

(١) الحصول على نماذج التركيب النحوي للتعبيرات الاسمية من ذخيرة تجريبية.

(٢) الاستعانة بالأخطاء في تشغيل طرق حساب عملية التعلم للحصول على القواعد التحويلية للتعبيرات الاسمية اعتمادًا على السياق.

(٣) الجمع بين كل من نماذج التركيب النحوي والقواعد التحويلية السياقية في تمييز التعبيرات الاسمية داخل النص.

وقد سبق أن قدمنا نتيجة التجربة للجزء الأول، وكان ذلك في النقطة (٥-٣) من الفصل الحالي، وفيما يلي نقدم نتيجة التجربة للجزء الثاني والثالث.

أولاً: تجربة الحصول على قواعد التحويل السياقية

يتم الاستعانة بالأخطاء في تشغيل طرق حساب عملية التعلُّم للحصول على القواعد التحويلية؛ حيث تم الحصول على إجمال ٣٨٠ قاعدة تحويل من ذخيرة تعلُّم يصل حجمها إلى حوالي خمسين ألف كلمة مع جعل القيمة المبدئية المُحَدَّدة سابقًا $T = 0$. وفيما يلي نورد عدد عشر قواعد تحويل هي الأكثر استخدامًا:

١- تحويل علامة الترميز المُرَشَّحة إلى علامة ترميز مؤكدة:

$when\ POS(p_{-1}) = QN.AND.POS(P_1) = 。$ (١)

مثال:

该/R公司/NG今年/T与/CM外商/NN签订/VN两/A项/QN]
承包/VNN合同/NG[。/。

"وقَّعَت تلك الشركة في العام الحالي عقدين للمقاولات مع رجال أعمال أجانب".

٢) تحويل علامة الترميز المُرَشَّحة إلى علامة ترميز مؤكدة:

---

(١) العلامة " 。" علامة ترقيم صينية تُستَخدَم في نهاية الجملة، وهي تساوي النقطة "." في اللغة العربية.

$$when\ POS(p_{-1}) = CM.AND.POS(P_{-2}) = BN$$

مثال:

……将/P个/R种/QN[反/H坦克/NG火力/NG]和/CM[防/H坦克/NG障碍物/NG]密切/A结合/VGN……

"... التأكيد على ضرورة الجمع بين قدرة الدبابات على المقاومة والدفاع...".

٣) تحويل علامة الترميز المُرَشَّحة إلى علامة ترميز مؤكدة:

$$when\ POS(p_{-1}) = ``.AND.POS(P_1) = "$$

مثال:

这/R种/QN语法/NG已经/D成为/VGN许多/MG立足/VGO于/P"/"[复杂/A特征/NG]"/"的/USDE"/"[合一/NG运算/VNN]"/"的/USDE[形式化/VNO方法/NN]的/USDE基础/NG。/。

"أصبح هذا النوع من القواعد النحوية أساس المنهجية الصورية التي تعتمد عليها الكثير من طرق الحساب التي تَستَمِد مقوماتها من الجمع بين الخصائص اللغوية المعقدة".

٤) تحويل علامة الترميز المُرَشَّحة إلى علامة ترميز مؤكدة:

$$when\ SENSE(p_{-1}) = Ja02$$

مثال:

这/R种/QN气候/NG叫做/VGN/Ja02[热带/NG雨林/NG气候/NG]……

"هذا النوع من الطقس يُطلَق عليه اسم طقس Ja٠٢ الاستوائي المُمطر...".

٥) تحويل علامة الترميز المُرَشَّحة إلى علامة ترميز مؤكدة:

$$when\ POS(p_{-1}) = P.AND.POS(P_1) = V$$

مثال:

在/P[上海/NG战役/NG]结束/VGO后/F，……

"بعد انتهاء معركة شنغهاي..."

٦) تحويل علامة الترميز المُرَشَّحة إلى علامة ترميز مؤكدة:

$$when\ SENSE(p_{-1}) = M$$

مثال:

许多/MG[无愧/VGO企业/NG]将/D转产/VGO，……

"العديد من المشروعات الخاسرة سوف تغير نشاطها...".

٧) تحويل علامة الترميز المُرَشَّحة إلى علامة ترميز مؤكدة:

$$when\ POS(p_{-1}) = M.AND.POS(P_1) = U$$

مثال:

许多/MG地方/NG分布/VN着/UT茂密/A的/USDE热带/NG雨林/NG，……

"العديد من المناطق مُعَرَّضة لأمطار استوائية...".

٨) تحويل علامة الترميز المُرَشَّحة إلى علامة ترميز مؤكدة:

$$when\ SENSE(p_{-1}) = Hc11.AND.SENSE(END(W)) = Dk14$$

مثال:

两/MJ国/NG政府/NG今天/T联合/D发表/VGN/Hc11建交/VGO公报/NG/Dk14，……

"أصدرت حكومتا الدولتين اليوم إعلانًا مشتركًا بإقامة علاقات دبلوماسية...".

٩) تحويل علامة الترميز المُرَشَّحة إلى علامة ترميز مؤكدة:

$$when\ POS(p_{-1}) = D.AND.POS(BEGIN(W)) = VGN$$

مثال:

两/MJ国/NG政府/NG今天/T发表/VGN[建交/VGO公报/NG]，……

"أصدرت حكومتا الدولتين اليوم إعلانًا بإقامة علاقات دبلوماسية...".

١٠) تحويل علامة الترميز المُرَشَّحة إلى علامة ترميز مؤكدة:

$$when\ SENSE(p_{-1}) = LeO2.AND.POS(p_1) = 。$$

مثال:

……组成/VGN/Le02[防/H步兵/NG火力/NG配系/NG]。/。

"... تكوين نظام مُدْمَج لمقاومة نيران المشاة".

من خلال الأمثلة السابقة، يمكننا أن نلاحظ مدى صحة الاتجاه إلى استنباط قواعد التحويل من خلال طريقة التعلُّم من الأخطاء. فلو نظرنا بشكل مستقل إلى إحدى القواعد، فمن المحتمل ألا تكون صحيحة بصورة كاملة (مثل: القاعدة الموضحة بالمثال رقم ٦)؛ إن خطوات التحويل الخاصة بتلك

القاعدة لا يتم الاستعانة بها تمامًا في بيئة التنشيط المحيطة بها. إلا أن كل قاعدة يتم النظر إليها على أنها صحيحة من حيث قدرتها على التطبيق والاستخدام، وأن نسبة الخطأ الضئيلة التي يمكن أن تنشأ من استخدام تلك القاعدة أو غيرها في الترميز يمكن تعويضها من خلال القواعد التالية لها (كما حدث في المثال رقم ٧). لذلك، فإن مجموعة القواعد بأكملها تكون مُرَتَّبة والقواعد التي تَرِدُ في البداية تكون أكثر عمومية، أما تلك التي ترد في مؤخرة القائمة فتكون أكثر خصوصية.

ثانيًا: تجربة تمييز التعبيرات الاسمية baseNP

تتمثل الخطوات التي تتبعها برمجية تمييز التعبيرات الاسمية baseNP فيما يلي:

١) الاعتماد على نماذج التراكيب النحوية الأساسية للتعبيرات الاسمية في عمل ترميز مبدئي للنص موضع الترميز.

٢) عمل تحويل لنتيجة الترميز في المرحلة السابقة من خلال تطبيق القواعد التحويلية واحدة تلو الأخرى.

٣) إذا استمر وجود لبس في الترميز في النتيجة النهائية، توضع علامتان أو أكثر أمام سلسلة الكلمات؛ مثل:

سلسلة الكلمات:

$$w_1/t_1 w_2/t_2 \ldots w_i/t_i \ldots w_j/t_j \ldots w_{N-1}/t_{N-1} w_N/t_N$$

الترميز رقم 1

الترميز رقم 2

ويتم الاحتفاظ بالترميز الأنسب وتُحذف العلامات الزائدة.

وحتى يمكن أن يتم عمل دراسة مقارِنة على عمليات تمييز التعبيرات الاسمية baseNP، قام جاو جوين بتقسيم التجربة إلى قسمين؛ حيث يختص القسم الأول بنماذج التراكيب النحوية الأساسية للتعبيرات الاسمية، ومن نتيجة الترميز في هذا القسم يمكننا أن نستنبط الخطوط الأساسية (base-line) لعملية تمييز التعبيرات الاسمية baseNP. ومن خلال القسم الأول يتم تنفيذ الخطوات ١، ٣ من التجربة السابقة. أما القسم الثاني من التجربة فيجمع بين كل من نماذج التراكيب النحوية الأساسية للتعبيرات الاسمية والقواعد التحويلية السياقية، بمعنى تنفيذ خطوات التجربة السابقة ١، ٢، ٣ على التوالي.

وتنقسم التجربتان بدورهما إلى جزأين، الأول هو الاختبار المغلق، والثاني هو الاختبار المفتوح، ويصل حجم النصوص التي يتم إدخالها في كل من التجربتين إلى عشرة آلاف كلمة. إلا أن النصوص الخاصة بالاختبار المغلق يتم اختيارها من ذخيرة التدريب، أما النصوص الخاصة بالاختبار المفتوح فيتم اختيارها من نصوص خارج نطاق ذخيرة التدريب.

وعند اختبار مؤشرات الكفاءة لبرمجية تمييـــز التعبيـــرات الاســـمية baseNP تم إقرار مؤشرين للقياس هما معدل الدقة P، ومعدل التغطية R، ويتم حسابهما من المعادلتين التاليتين:

معدل الدقة:

$$P = \frac{a}{b} \times 100\%$$

معادلة (٥-٢)

معدل التغطية:

$$R = \frac{a}{c} \times 100\%$$

معادلة (٥-٣)

حيث تُشير a في المعادلتين إلى عدد التعبيـــرات الاســـمية baseNP التي ميزتها البرمجية بشكل صحيح في النص موضع المعالجة، وتُشير b إلى إجمالي عدد سلاسل الكلمات التي تم الحكم أنها تعبيرات اسمية، أما c فتشير إلى عدد التعبيرات الاسمية الموجودة بالفعل داخل النص موضع المعالجـــة. ويشير الجدول ٥-١٣ إلى نتيجة اختبار التجربتين:

جدول (٥-١٣): مقارنة بين الطريقتين المستخدمتين في تمييز التعبيرات الاسمية

نوع الاختبار	الطريقة التي تستخدم النماذج الأساسية في التركيب النحوي		الطريقة التي تعتمد على الجمع بين النماذج الأساسية في التركيب النحوي والقواعد التحويلية	
	نسبة الدقة	نسبة التغطية	نسبة الدقة	نسبة التغطية
الاختبار المغلق	٧٢,٩%	٧٧,٥%	٩١,١%	٩٣,٢%
الاختبار المفتوح	٧٢,٧%	٧٨,٣%	٨٧,٣%	٩١,٨%

من خلال مقارنة نتيجة التجربتين السابقتين، يمكننا ملاحظة أن نتائج الطريقة المختلطة التي جمعت بين النماذج الأساسية في التركيب النحوي والقواعد التحويلية قد تَفَوَّقَت بوضوح على طريقة النماذج الأساسية في التركيب النحوي.

وفيما يلي نقدم جزءًا من نتيجة الترميز التي أُجْرِيَت على نص الاختبار المفتوح (الأجزاء المكتوبة بالبنط السميك؛ التي أسفلها خط تشير إلى خطأ في الترميز):

[干部/NG工作/NG]是/VY中国人民解放军/NG依据/P中国共产党/NG的/USDE[干部/NG路线/NG]和/CM[政策/NG管理/VNN军官/NG]和/CM[文职/NG干部/NG]的/USDE工作/NG。/。[机构/NG干部/NG工作/NG]，/，原来/D是/VY[中国人民解放军/NG建设/VNN]的/USDE[重要/A内容/NG]。/。根据/P1929年/T[古田/NPL会议/NG决议/NG]的/USDE规定/NG，/，[工农/NG红军/NG]的/USDE[军事/NG干部/NG]由/P[军事/NG系统/NG]管理/VNN，/，其/R[具体/A工作/NG]，/，由/P[司令/NG机关/NG]的/USDE[队列/NG部门/NG]和/CM[政治/NG机关/NG]的/USDE[组织/NG部门/NG]负责/VGV，/，1973年/T以后/F，/，干部/NG的/USDE任免/VNN、/、调配/VNN由/P各/R级/NG[军政/NG委员会/NG]按/P[任免/VNN期限/NG]讨论/VNN决定/VGN，/，有的/R部队/NG还/D在/P队列/NG和/CM[组织/NG部门/NG]内/F成立/VGN了/UT干部科/NG。/。

الفصل الثالث

نماذج إزالة اللبس الدلالي للكلمات الصينية

اعتمادًا على الفراغ الدلالي للكلمات داخل التركيب

إن المقصود بإزالة اللبس الدلالي للكلمات (word sense disambiguation) هو معالجة الكلمات متعددة الدلالة بحيث يتم وضع كود دلالي لها وفقًا للمحيط السياقي الذي ترد فيه. وهذا الكود الدلالي قد يكون عبارة عن الرقم الذي يشير إلى معنى تلك الكلمة داخل أحد المعاجم العامة، وقد يكون عبارة عن كود التصنيف الدلالي لتلك الكلمة داخل أحد المعاجم الدلالية، أو الكلمة المقابلة لتلك الكلمة في معجم للترجمة؛ كما يمكن أن يكون شرحًا لتلك الكلمة داخل أحد المعاجم المفهومية. وقد ساد لفترة طويلة اعتقاد أن موضوع إزالة اللبس الدلالي عن الكلمات هو أحد الموضوعات الصعبة في مجال معالجة اللغات الطبيعية. فقبل تسعينيات القرن الماضي كانت الدراسات المتعلقة بإزالة اللبس الدلالي تعتمد على منهجيات الذكاء الاصطناعي بصورة أساسية، وكانت الصعوبة في ذلك تكمن في حاجة هذه المنهجيات إلى العنصر البشري من أجل صياغة عدد هائل من قواعد إزالة اللبس، وسلبيات ذلك ليست ضئيلة. ليس فقط من حيث نسبة التغطية، بل إن هادر تلك المنهجيات كبير جدًا، وهذا ما أُطلق عليه "عنق الزجاجة" في مجال

استخلاص المعارف من النصوص. وبعد تسعينيات القرن العشرين، دخلت الدراسات المتعلقة بإزالة اللبس الدلالي عن الكلمات عصرًا جديدًا يتخذ من منهجية الذخائر اللغوية منطلَقًا أساسيًّا؛ وذلك بسبب ظهور المعاجم الآلية والذخائر اللغوية هائلة الحجم.

تعتمد منهجيات إزالة اللبس الدلالي عن الكلمات المُحتواة في معاجم إلكترونية بصورة كبيرة على النصوص الشارحة للكلمات داخل المعاجم العامة؛ حيث يتحقق غرض إزالة اللبس الدلالي لكلمة بمجرد حساب درجة تكرار النص الشارح لمعنى محدد من المعاني المتعددة لتلك الكلمة داخل النص موضع المعالجة؛ وذلك مثلما فعل العالمان ليسك (**Lesk**) وويلكز (**Wilks**)، والذي قَدَّم كل منهما منهجيته الخاصة لإزالة اللبس الدلالي[104,103].

إلا أنه عندما يكون النص الشارح قصيرًا، كأن يقتصر على سرد الكلمة المضادة أو المترادفة مع الكلمة موضع التعامل، فمن الصعب العثور على معلومات تكرارية لمثل هذا النص الشارح داخل النص؛ ومن ثَمَّ يؤثر ذلك في نتيجة إزالة اللبس. وهناك طرق أخرى تعتمد على إزالة اللبس من خلال استخدام معجم التصنيفات الدلالية. وفي هذا المجال تُعتَبر الطريقة التي قَدَّمَها ياروسكي (**Yarowsky**) هي الطريقة الأكثر تمثيلاً لهذا النوع من منهجيات إزالة اللبس الدلالي[105]. وقد اعتمدت تلك الطريقة عند حساب الكلمات ذات البروز الدلالي (**salient words**) داخل النص على توزيع متوسط عدد مرات ظهور الكلمة متعددة الدلالة على كل تصنيف دلالي مقابل لتلك الكلمة، الأمر الذي تسبب في وجود ضجيج إحصائي، فضلاً عن

محدودية المادة اللغوية المستخدمة في رصد الكلمات ذات البروز الـدلالي؛ ومن ثَمَّ فقد كانت نسبة التغطية لهذه الطريقة محدودة جدًّا.

إن منهجية إزالة اللبس الدلالي اعتمادًا على ذخيرة لغوية؛ التي قـدمها كل من ياروسكي (Yarowsky)، وبروس (Bruce)، وغيرهمـا[107,106] تعتمد بصورة كبيرة على الترميز الذي يتم من خلال العنصر البشري علـى ذخيرة تدريبية، وهذا النوع من الترميز مُكلِّف من حيث الوقت والمال، هـذا بالإضافة إلى وجود مشكلة في النتائج الإحصائية تتمثل في خلخلة البيانـات وعدم ترابطها؛ لذلك يسعى عدد من العلماء إلى بحث إمكانية التوصل إلـى منهجيات لاستخراج المعلومات من النصوص اعتمادًا علـى إزالـة اللـبس الدلالي غير الموجه (unsupervised). إلا أن هذه المنهجيات ما زالت تقف حتى اليوم عند حد تجارب على نطاق ضيق لا يتعدى بضع كلمات أو بضع عشر كلمة متعددة الدلالة. وقد طرحت لي جوان تزي من جامعة تشين خوا نموذجًا لإزالة اللبس الدلالي اعتمادًا على الفراغ الـدلالي للكلمـات داخـل التركيب. ونظرًا إلى أن كل مجموعة مترادفات في معجم (غابة المترادفات) يقابلها كود واحد يمثل كل حقل دلالي، هذا بالإضـافة إلـى أن المجموعـة الواحدة من المترادفات تحتوي دائمًا عددًا قليلاً من الكلمات متعددة الدلالـة، وكمية كبيرة من الكلمات أحادية الدلالة. ولذلك فمن الممكن أن نستخرج من ذخيرة كبيرة الحجم مجموعة الكلمات الحقيقية التي تسبق كلمة مفـردة مـن الكلمات التي تُكَوِّن مجموعة كلمات مترادفة وتلحقها، وإنشاء برمجية تقطيع آلية تتخذ من مجموعة الكلمات تلك كودًا للتصنيف الدلالي. ونظرًا إلـى أن

هذا النوع من التعلُّم الآلي في اكتساب المعلومات الخاصة بإزالة اللبس الدلالي عن الكلمات غير مُوَجَّه؛ ومن ثَمَّ، فإن بإمكانه تقليل الكثير من الأعباء المالية التي يفرضها تدخل العنصر البشري في عمل الترميز الدلالي للكلمات داخل الذخائر اللغوية. وقد أثبتت التجارب أن أنماط إزالة اللبس الدلالي تلك تتميز بنسبة دقة عالية في إزالة اللبس الدلالي عن الكلمات، هذا بالإضافة إلى أنها تتميز بإمكانية الانتقال في التطبيق عبر أنواع النصوص المختلفة. وفيما يلي نقدم تعريفًا بالبحث الذي قامت به الباحثة لي جوان تزي.

١- تعريف بمعجم (غابة المترادفات)

اعتمد مؤلف معجم (غابة المترادفات)[108] (يُطْلَق عليه اختصارًا (غابة الكلمات) في أثناء وضع التصنيفات الدلالية للكلمات على الكلمة كعنصر أساسي، مع الاستعانة برموز التصنيف النحوي لها، بالإضافة إلى الاهتمام الشديد بالكثافة النسبية لموضوع الحديث. وهذا المعجم الخاص بالتصنيفات الدلالية قد قَسَّم الكلمات من حيث المعنى إلى ثلاثة مستويات: مستوى التصنيف الأكبر، والأوسط والأصغر، وفي هذا الإطار قام بتقسيم الكلمات إلى عدد ١٢ تصنيفًا كبيرًا، وعدد ٩٤ تصنيفًا وسطًا، وعدد ١٤٢٨ تصنيفًا صغيرًا، ويضم التصنيف الصغير عددًا من الكلمات كعناوين يندرج في إطارها مجموعات الكلمات المترادفة، ويصل عدد تلك العناوين إلى ٣٩٢٥ عنوانًا.

يَستَخدِم معجم (غابة المترادفات) الحرف اللاتيني الأول في شكله الكبير ككود للتصنيف الأكبر، يليه مباشرة الحرف الثاني في شكله الصغير ككود للتصنيف الأوسط، أما الكود الثالث والرابع في التصنيف الدلالي فتُستَخدَم له الأرقام العربية؛ حيث يشير كل منها إلى كود التصنيف الأصغر. والعناوين التي تندرج تحت التصنيف الأصغر يُستَخدَم لها أيضًا الأرقام العربية. على سبيل المثال الكلمة "觉悟" "إدراك"، كود التصنيف الدلالي لها هو "Ga15"؛ حيث يشير الحرف "G" في هذا الكود إلى التصنيف الأكبر ويضم الكلمات التي تُعبِّر عن "النشاط النفسي" أما الحرفان "Ga" فيشيران إلى التصنيف الأوسط ويضم الكلمات التي تُعبِّر عن "الحالة النفسية"، أما كود التصنيف الأصغر فهو "Ga15"، وهكذا يتم التعبير عن ذلك في المعجم كما يلي:

Ga15 醒悟　懂事
醒悟觉悟省悟憬悟觉醒清醒如梦初醒大梦初醒……
懂事　记事儿　开窍　通窍

أي أن الكود"Ga15" يضم تحتَه كلمتان تعبر كل منهما عن عنوان لمجموعة من المترادفات هما "醒悟" "انتباه"، و"董事" "فهم". لذلك فإن الكود الدلالي الكامل لكلمة "觉悟" "إدراك" هو **Ga1501** وهذا لأن هذه الكلمة وقعت في الموضع الثاني في قائمة المترادفات.

ونظرًا إلى أن المداخل الدلالية في معجم (غابة المترادفات) تتخذ من الكلمات المفردة وحدة لها، فإن الكلمات متعددة الدلالة تأخذ أكوادًا مختلفة وفقًا لمعانيها. على سبيل المثال الكلمة "材料" لها ثلاثة مداخل دلالية في معجم (غابة المترادفات) كما يلي:

(١) مادة تدخل في صناعة منتج.

(٢) مادة مرجعية تدخل في تأليف كتاب أو مواد علمية يتم تقديمها بغرض الاطلاع.

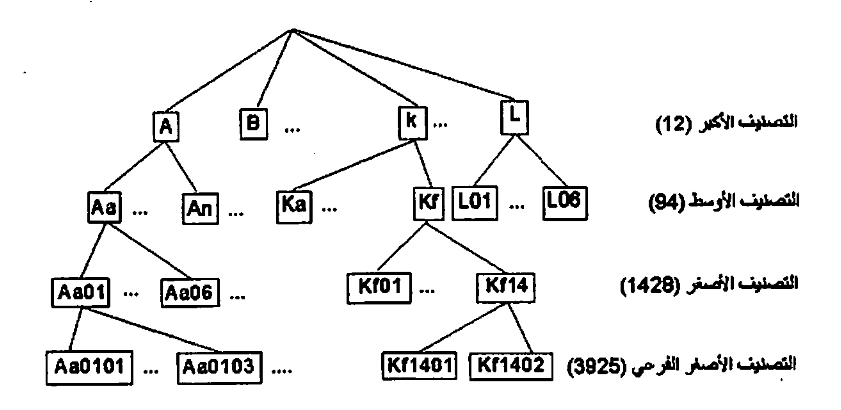
(٣) تُستَخدَم للكناية عن بعض الأشخاص الذين يصلحون لأداء مهمة معينة.

وهذه المعاني يقابلها الأكواد التالية على التوالي: "**Ba06**"، "**Dk17**"، "**A103**".

وفي أثناء إزالة اللبس الدلالي عن كلمة "材料" يتم وضع الكود المناسب لها وفقًا للسياق الذي يظهر معها داخل النص.

وكما سبق أن أشرنا نجد أن منظومة ترميز التصنيفات الدلالية داخل معجم (غابة المترادفات) عبارة عن تركيب شجري كما يتضح من الشكل ٥-٤.

شكل (٥-٤): التركيب الشجري لمنظومة أكواد التصنيفات الدلالية بمعجم ((غابة المترادفات))



وفي الحقيقة، إن الكلمات التي ضمها معجم (غابة المترادفات) يضم عدد ٥٠١٥٤ مادة موزعة ما بين الكلمات بشكل أساسي، بالإضافة إلى بعض التعبيرات اللغوية والأمثال والكلمات المأثورة. ويوضح الجدول رقم ٥-١٤ توزيع الكلمات متعددة الدلالة داخل معجم (غابة المترادفات)؛ التي يصل عددها الإجمالي إلى ٧٤٣٠ كلمة وتشغل نسبة ١٤,٨%. والجدير بالذكر أن ما يقرب من نصف عدد الكلمات أحادية المقطع؛ التي يصل عددها إلى ٣٧٧٤ كلمة بما يمثل ٧,٥٢% من العدد الكلي للكلمات التي ضمها المعجم (عدد ١٨٠١ كلمة)، هي كلمات متعددة الدلالة؛ وبالمقارنة نجد أنه من بين عدد ٤٦٣٨٠ كلمة متعددة المقطع، يوجد نسبة ١٢,١% كلمة متعددة الدلالة.

جدول (٥-١٤): توزيع الكلمات متعددة الدلالة في معجم (غابة المترادفات)

	كلمات أحادية المقطع		كلمات متعددة المقطع		إجمالي عدد الكلمات
	عدد الكلمات	النسبة المئوية	عدد الكلمات	النسبة المئوية	
كلمات أحادية المعنى	١٩٧٣	%٥٢,٣	٤٠٧٥١	%٨٧,٩	٤٢٧٢٤
كلمات متعددة المعنى	١٨٠١	%٤٧,٧	٥٦٢٩	%٢١,١	٧٤٣٠
الإجمالي	٣٧٧٤	%١٠٠	٤٦٣٨٠	%١٠٠	٥٠١٥٤

إن درجة الصعوبة أو السهولة في مهمة إزالة اللبس الدلالي عن الكلمات يُمْكِن أن تظهر من خلال استكشاف الذخيرة. فعلى سبيل المثال، قد نجد أن ما يقرب من ٤٢% من إجمالي عدد كلمات الذخيرة عبارة عن كلمات لها أكثر من معنى. وكما سبق أن أشرنا، فإن التصنيفات الدلالية في معجم (غابة المترادفات) قد استعانت بالتصنيفات النحوية للكلمات، على سبيل المثال نجد أن التصنيفات الكبرى التي تشير إليها الحروف من **A~D** تندرج تحت تصنيف الأسماء، أما الأرقام والكلمات الكمية فتندرج تحت التصنيف الأوسط وتأخذ الحرفين **Dn**، والحرف **E** يشير إلى الصفات، أما الحروف من **F~J** فتشير إلى تصنيف الأفعال. لذلك فإنه بالنسبة إلى نص قد سبق تمييز حدود الكلمات به وترميزها من حيث التصنيف النحوي، فإنه يمكن تمييز معنى عدد غير قليل من الكلمات متعددة الدلالة الموجودة به وفقًا للتصنيف النحوي لتلك الكلمات. وتشير البيانات الإحصائية أنه بعد إزالة اللبس عن التصنيف النحوي للكلمات، انخفض عدد الكلمات متعددة الدلالة داخل الذخيرة من ٤٢% إلى ٢٤%، وقد قلت نسبة تواجدها داخل الذخيرة بمقدار ٤٣%.

٢- التعبير عن فراغ المُتَّجَهات في معجم ((غابة المترادفات))

"يمكنك معرفة معنى الكلمة من خلال الكلمات المصاحبة لها"، هذا ما قاله عالم اللغة فيرث (Firth)[109] عند وصفه لطريقة تمييز معنى الكلمة. ويُقْصَد بذلك أن معنى الكلمة لا يمكن الاستدلال عليه إلا من خلال الاستخدام. وعلى هذا، فإن استطلاع السياق المصاحب لكلمة ما في كل مرة من مرات ظهورها داخل ذخيرة لغوية، يُمَكِّنُنا من الحصول على معاني تلك الكلمة من خلال العلاقات التصاحبية بينها وبين الكلمات الأخرى. ولا يقتصر الأمر على أن لكل كلمة مجموعة العلاقات التصاحبية التي ترد معها فحسب، بل إن الكلمة الواحدة تختلف مجموعة التصاحبات التي ترد معها كلما اختلف معناها.

ونظرًا إلى أن معنى كلمة ما يمكن وصفه من خلال مجموعة الكلمات التي تظهر معها (باختصار علاقات التصاحب)، لذلك فمن الممكن استخدام مُتَّجَه متعدد الأبعاد في التعبير عن معنى محدد للكلمة. وقد قامت لي جوان تزى بتعريف هذا النوع من المُتَّجَهات بأنه مُتَّجَه دلالة الكلمة. وبمزيد من التفصيل فإن مُتَّجَه دلالة الكلمة يتكون من مجموعة من العناصر، وكل عنصر من هذه العناصر يمثله كلمة من الكلمات الحقيقية التي تتصاحب مع الكلمة متعددة الدلالة، ويمثل هذا المُتَّجه أحد أبعاد الفراغ الدلالي للكلمة.

وعند التفكير في متطلبات تنفيذ هذا المشروع، من المفضل تحديد "السياق" المصاحب لدلالة الكلمة على أنه مجموعة الكلمات الحقيقية التي تَرِد قبل هذه الكلمة أو بَعدَها، التي يُرْمَز لموضعها بالرمز d، وعلى هذا فإن $\pm d$ يُطلَق عليها نافذة الاستقراء التي تُستَخدم في استطلاع التصاحبات اللغوية.

ونظرًا إلى أن هذه المجموعة من الكلمات الحقيقية تختلف في قدرتها على الظهور في إطار الفلك الدلالي للكلمة موضع الدراسة، فإن هناك ضرورة لاستخدام وزن (weight) مُحدَّد لكل كلمة حتى يمكن تمييز قدرة كل واحدة منها. وقد عَرَّفَت لي جوان تزي وزن الكلمات الحقيقية التي تمثل المُتَّجَهات الدلالية للكلمة على أنه احتمال تصاحب الظهور بين الكلمات الحقيقية x_i ودلالة كلمة معينة s ويرمز لهذا الوزن بالرمز $p(s,x_i)$. ويبدو من ذلك أن وزن الكلمات المصاحبة لدلالة الكلمة $p(s,x_i)$ يمكن تقدير قيمته من خلال إجراء بعض العمليات الإحصائية على ذخيرة لغوية.

ومن خلال التعديل السابق نجد أن كل عنصر من العناصر المُكَوِّنَة لمُتَّجَه دلالة الكلمة V يمكن الإشارة إليه باستخدام احتمال تصاحب الظهور $p(s,x_i)$، أي أن $Vx_i = P(s,x_i)$. لذلك، ففي حقيقة الأمر أن المُتَّجَه الدلالي للكلمة يساوي قيمة حقيقية لمُتَّجَه متعدد الأبعاد، وهذا المُتَّجَه الدلالي عبارة عن مجموعة المُتَّجَهات التي تُكَوِّن الفراغ الدلالي للكلمة، ويُطلَق عليه أيضًا اسم فراغ المُتَّجَه الحقيقي متعدد الأبعاد.

هذا ويعتمد الوصف السابق لمعنى الكلمة على مرجعية الفرضين التاليين:

[الفرض الأول] إذا تساوت دلالة كلمتين، فإن ذلك يؤدي إلى تساوي الكلمات التي تتصاحب معهما في السياق النصي. وإذا استخدمنا المُتَّجَه الدلالي في التعبير عن السياق المصاحِب لهاتين الكلمتين، فسوف يكون هناك تقارب في المسافة بين الفراغ الدلالي لهما.

[الفرض الثاني] هناك عدد من الكلمات تتشابه أو تتقارب في المعنى، وهذه الكلمات تظهر كأنها ائتلاف من النقاط المتجاورة من حيث قيمة الفراغ الدلالي الذي يمثلها.

ومدى موثوقية الفرض الثاني يمكن إثباتها من خلال صفة التَّعَنْقُد التي تتميز بها دلالات الكلمات، والغرض من ذلك هو اختبار مدى التطابق بين نظام التصنيف الدلالي المُتَّبَع في معجم (غابة المفردات) وبين مجموعات الكلمات المترادفة (متقاربة المعنى) التي يتم الحصول عليها من خلال الفرض الثاني، والذي يمثل صفة التعنقد للكلمات المترادفة والمتقاربة في المعنى. وقد صمَّمَت لي جوان تزي تجربتها كما يلي: يتم اختيار أي مجموعتين من الكلمات A، B من معجم (غابة المترادفات)؛ حيث تمثل كل منهما فئة دلالية أصغر، وبشرط أن تتساويا من حيث التصنيف النحوي، وعلى فرض أن كلا من C_A، و C_B تُشيران إلى قائمة الكلمات الكاملة التي تندرج تحت المجموعة A، و B على التوالي، أي أن:

$$C_A = \{WA_1, WA_2, ..., WA_m\}$$

$$C_B = \{WB_1, WB_2, ..., WB_n\}$$

حيث تشير $WA_i(i=1,...,m)$ إلى كلمة أحادية المعنى في المجموعة A وتشير $WB_j(j=1,...,n)$ إلى كلمة أحادية المعنى في المجموعة B. ووفقًا لمبادئ تكوين المُتَّجه الدلالي للكلمات، يمكننا الحصول على المُتَّجَه الدلالي لأي كلمة $V(W)$ من الكلمات سالفة الذكر داخل ذخيرة كبيرة الحجم. ثم يتم

إعادة تصنيف جميع الكلمات التي تقع في إطار مجموعتي الكلمات $C=C_A \cup C_B$ من حيث مدى تقارب المسافة بين مُتَّجَهاتِها أو تباعدها، وقد أظْهَرَت نتيجة التعنقُد أن مجموعتي الكلمات C_1، و C_2 تحتويان $C=C_1 \cup C_2$، بالإضافة إلى أن $C_1 \cap C_2 = \Phi$، وإذا كانت كلٌّ من C_1، و C_2 متطابقتان لدرجة معينة مع كل من C_A، و C_B فإن هذا يعني أن الفرض الثاني فرضًا صحيحًا.

عند استخدام طريقة حساب المسافات الأقْصَر السابق ذكرها؛ التي تتخذ الاتجاه من أسفل إلى أعلى عند إجراء عملية جمع التصنيفات، تكون البداية أن نضع الكلمات التي تتدرج في القائمة C؛ التي يزيد عدد مرات تكرارها عن مائة مرة في المجموعة C_1 والمجموعة C_2، ثم نقوم بتكرار عملية جمع التصنيفات بالطريقة نفسها على الكلمات التي تكررت بعدد مرات أقل.

وصل نطاق الذخيرة المُستَخْدَمة في تجربة جمع التصنيفات إلى ٧٢ ميجا بايت. ويحتوي الجدول ٥-١٥ أزواج أكواد التصنيف الدلالي المستخدمة في التجربة، بالإضافة إلى معلومات عن معدل تكرار تلك الأكواد داخل الذخيرة. ويعرض الجدول ٥-١٦ نتيجة التجربة. وقد تم حساب معدل التوافق بين كل من التصنيف الدلالي للكلمات القائم على المُتَّجَهات الدلالية، وأكواد التصنيف الدلالي المستخدمة في معجم (غابة المترادفات) من خلال المعادلة التالية:

معدل التوافق=	عدد الكلمات المشتركة بين نتيجة التصنيف باستخدام المُتَّجَهات الدلالية وبين أكواد معجم ((غابة المترادفات)) ――――――――――― عدد الكلمات الكلي في القائمة C

جدول (٥-١٥): نماذج من أزواج الأكـواد الدلاليـة المستخدمة في معجم ((غابة المترادفات))

وعدد مرات ظهورها في الذخيرة

أزواج أكواد التصنيف الدلالي	عدد الكلمات أحادية الدلالة	العدد الكلي للكلمات	أكبر من ١٠٠ مرة	أكبر من ٥٠ مرة	أكبر من ١٠ مرات
Hc11/Hc03	18/17	6538/6005	7/6	9/9	11/13
Ba06/Da19	16/16	3954/3415	4/5	5/6	8/10
Hc11/Hi03	18/17	6165/6005	5/6	7/9	13/13
Aa03/Ae07	20/15	6735/6800	4/6	4/6	9/10
Di10/Di08	28/27	11531/12017	8/8	9/11	3/20
Ed29/Ed11	17/15	4054/3534	3/2	6/3	10/8
Ed16/Ef08	17/14	2656/2599	3/1	4/2	8/5
Gb15/Hj20	6/7	2003/2303	2/4	3/4	5/6

حيث تعني "أزواج أكواد التصنيف الدلالي" في الجدول السابق زوج التصنيف الدلالي الذي يتم اختياره من معجم (غابة المترادفات)، أما "عدد المفردات أحادية الدلالة" فيعني عدد الكلمات التي لها معنى واحد داخل قائمة المترادفات التي يمثلها كل كود من أكواد التصنيف الدلالي، أما "العدد الكلي للكلمات" فيعني العدد التراكمي الذي يمثل مرات ظهور هذه الكلمات أحادية المعنى داخل الذخيرة، أما "أكبر من ١٠٠" فتعني عدد الكلمات أحادية المعنى داخل كل تصنيف دلالي؛ التي يتجاوز عدد مرات ظهورها مائة مرة في الذخيرة، والباقي من عناوين الأعمدة على القياس نفسه.

جدول (٥-١٦): معدل التوافق بين نتيجة التصنيف وأكواد معجم (غابة المترادفات)

أزواج أكواد التصنيف الدلالي	أكبر من ١٠٠ مرة	أكبر من ٥٠ مرة	أكبر من ١٠ مرات	متوسط معدل التوافق
Ba06/Da19	%١٠٠	%١٠٠	%١٠٠	%١٠٠
Aa03/Ae07	%٩٠,٠	%٩٠,٠	%٨٤,٤	%٨٢,٥
Di10/Di08	%٨٧,٥	%٨٥,٠	%٧٥,٨	%٧٥,٥
Gb15/Hj20	%١٠٠	%٨٥,٨	%٨١,٨	%٨٤,٦
He11/Hi03	%٩٠,٩	%٩٠,٩	%٧٦,٤	%٦٩,٣
Hc11/Hc03	%١٠٠	%١٠٠	%٩١,٧	%٧٧,١
Ed16/Ef08	%١٠٠	%١٠٠	%٨٣,٣	%٨١,٣
Ed29/Ef11	%١٠٠	%١٠٠	%٨٤,٦	%٧٨,٦
متوسط معدل التوافق	%٩٦,١	%٩٠,٤	%٨٤,٨	%٨١,١

وتشير نتيجة التجربة إلى أنه:

(١) إذا نظرنا إلى التصنيف الدلالي للكلمات أحادية المعنى التي يزيد عدد مرات ورودها في الذخيرة عن ١٠٠ مرة وفقًا لقرب المسافة بين المُتَّجَهات الدلالية لهذه الكلمات أو بُعدها، نجد أن ما يزيد عن ٩٠% من هذه الكلمات يتوافق تمامًا مع معجم (غابة المترادفات)؛ حيث وصل متوسط معدل التوافق إلى ٩٦,١%. أما الكلمات التي تجاوز عدد مرات ظهورها ٥٠ مرة، فقد كان من بينها أكثر من ٨٢% يتوافق مع معجم (غابة المترادفات)، وقد كان متوسط معدل التوافق لهذه الفئة ٩٠,٤%. وهذا يعكس مصداقية الفرض الثاني.

(٢) إذا نظرنا إلى متوسط معدل التوافق بين نتيجة التصنيف الآلية وتصنيف معجم (غابة المترادفات)، نجد أن الكلمات التي لها معدل تكرار أعلى تَتَفَوّق بوضوح على الكلمات التي لها معدل تكرار أقل. والسبب في ذلك يرجع إلى أنه كلما زاد عدد مرات ظهور الكلمة، أصبحت البيانات الإحصائية أكثر مصداقية، فضلاً عن اقتراب مؤشرات المُتَّجَهات الدلالية من الواقع، وبالطبع يزداد متوسط معدل التوافق مع معجم (غابة المترادفات).

(٣) وبصورة عامة، فإن نتيجة التصنيف الناتجة عن أزواج الأكواد التي تنتمي إلى تصنيفات دلالية كبرى مختلفة تكون أفضل من نتيجة التصنيف لأزواج من الأكواد تنتمي إلى تصنيفات دلالية كبرى متشابهة، على سبيل المثال: **Ba06/Da19** أفضل من **Aa03/Ae07**. وهذا يدل على أنه كلما زاد التباين بين أكواد التصنيفات الدلالية في معجم (غابة المترادفات)، بَعُدَت المسافة بين الفراغ الدلالي للمتجهات الدلالية التي تمثل هذه الأكواد؛ ولذلك فمن السهل الحفاظ على مقدار من التوافق بين نتيجة التصنيف الآلية وبين تصنيف معجم (غابة المترادفات). وما ينبغي التأكيد عليه هو أن المرجعية التي اعتمد عليها معجم (غابة المرادفات) في وضع أكواد التصنيفات الدلالية التي تصنف مجموعات الكلمات المترادفة هي الرؤية الذاتية أو الحس اللغوي لعلماء اللغة، أما تكوين المُتَّجَهات الدلالية للكلمات فتعتمد على تصاحب الظهور بين

الكلمات داخل النصوص؛ بالإضافة إلى أن ذلك يتم من خلال استقراء ذخيرة نصوص ذات نطاق واسع، وهذا يشير إلى أن المنهجيتين مختلفتين بشكل كلي. إلا أن نتيجة التصنيف الدلالي التي توصل إليها الحاسب الآلي عن طريق قياس المسافة بين المُتَّجَهات الدلالية للكلمات كانت متطابقة إلى حدٍّ كبير مع نتيجة التصنيف الدلالي لمعجم (غابة المترادفات). وهذه الحقيقة تُشِير إلى منطقية الفرضين اللذين قدمتهما لي جوان تزي عن الوصف الدلالي للكلمات، هذا بالإضافة إلى أن الحس اللغوي لعلماء اللغة يمكن قياسه عند حدود معينة.

وخلاصة ما سبق أن أية مجموعة من الكلمات المترادفة دائمًا ما تحتوي عددًا من الكلمات أحادية الدلالة، وأن استخراج الكلمات الحقيقية التي تتصاحب مع هذه الكلمات أحادية الدلالة في ذخيرة واسعة النطاق أمرٌ يمكن تحقيقه آليًا بعيدًا عن تدخل العنصر البشري بصورة كاملة. إضافة إلى ذلك ووفقًا للفرض الثاني، فإن مجموعة الكلمات المترادفة (أي تلك التي يكون لها كود دلالي واحد) دائمًا ما يمكن استخدام مُتَّجَه دلالي واحد يشير إليها. وهذا المُتَّجَه الدلالي هو مركز المُتَّجَهات التي توضع لجميع الكلمات أحادية الدلالة في مجموعة الكلمات المترادفة.

إن كل مُتَّجَه فرعي في مُتَّجَه التصنيف الدلالي غالبًا ما يُمَثَّل بالكلمة الحقيقية المصاحِبَة؛ التي يُشار إليها بالرمز $x_i(i=1,...,n)$ ؛ حيث تشير n إلى عدد الأبعاد التي تُمَثِّل فراغ المُتَّجَه. فعلى فرض أن A تشير إلى الكلمات

أحادية الدلالة الموجودة في مجموعة كلمات مترادفة، فإن القيمة الموجودة في المُتَّجَه الفرعي x_i يمكن حسابها من خلال قيمة المتوسط الحسابي لمجموعة احتمالات $p(W, x_i)$ بين جميع الكلمات أحادية الدلالة والكلمة x_i، كما يلي:

$$V_{x_i} = \frac{1}{|A|} \sum_{W \in A} P(w, x_i)$$

حيث تشير $|A|$ إلى إجمالي عدد الكلمات في قائمة الكلمات أحادية الدلالة A، أما w فتشير إلى أية كلمة في القائمة A.

وقد قامت لي جوان تزي بتكوين مُتَّجهات للتصنيفات الدلالية الصغرى في معجم ((غابة المترادفات)) من خلال ذخيرة حجمها ٧٢ ميجا بايت مُكَوَّنَة من نصوص جريدة الشعب اليومية، وقد قامت بتكوين فراغ دلالي من هذه المُتَّجَهات الدلالية. وقد كانت المسافة قصيرة بين المُتَّجَهات الدلالية للكلمات المتقاربة في المعنى وسط هذا الفراغ الدلالي. ومن ثم، تَكَوَّنَت متجهات دلالية تُعَبِّر عن تلك الكلمات متقاربة المعنى؛ بالإضافة إلى أن هذه المُتَّجَهات الدلالية يمكن ضمها معًا لتكوين مُتَّجَه يُمثل طبقة دلالية معينة، وهكذا يتم تكوين فراغ دلالي للكلمات ذات تركيب متدرج كما يتضح من الشكل ٤-٥. وقيمة هذه الدراسة تكمن في أنها تُقَدِّم نوعًا من الصياغة المعلوماتية لكيفية إزالة اللبس الدلالي بطريقة حاسوبية، وهذه المعلومات الخاصة بإزالة اللبس من الممكن استخدام الحاسب الآلي في الحصول عليها آليًا من ذخيرة لغوية ذات نطاق متسع. ومن ثَمَّ، يمكن تجنب الأعباء الهائلة التي كانت تُلْقَى على عاتق العنصر البشري في الماضي لعمل الترميز الدلالي للذخائر اللغوية.

إن الاعتماد على هذا النموذج اللغوي في إزالة اللبس الدلالي عن الكلمات قد نشأ عنه اختصار عمليات تمييز معنى الكلمة متعددة الدلالة إلى خطوتين: الخطوة الأولى تتمثل في تكوين مُتَّجَه للفراغ الدلالي للكلمة متعددة الدلالة وفقًا للسياق الذي تَرِد فيه (عادة ما يكون هذا السياق جملة)، ثم يتم البحث عن المُتَّجَهات الدلالية المقابلة لهذه الكلمة متعددة الدلالة انطلاقًا من الفراغ الدلالي لهذه الكلمة. ويتم الحُكْم النهائي على معنى الكلمة في السياق الحالي من خلال تحديد أقرب مُتَّجَه دلالي لمُتَّجَه هذه الكلمة.

٣- نموذج إزالة اللبس الدلالي عن الكلمات اعتمادًا على الفراغ الدلالي للكلمة داخل التركيب

يتكون هذا النموذج الذي يُستَخْدَم في إزالة اللبس الدلالي عن الكلمات انطلاقًا من الفراغ الدلالي للتركيب من العناصر المُوَضَّحة بالشكل ٥-٥. وفيما يلي نشرح وظيفة كل عنصر من هذه العناصر:

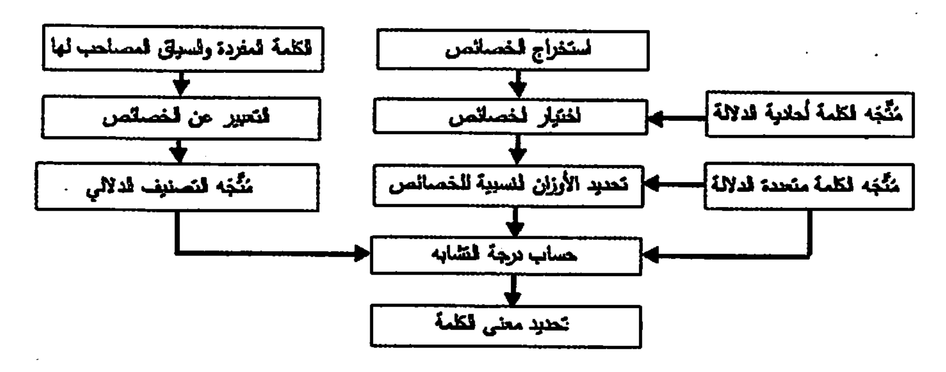
(أ) استخراج الخصائص: وذلك من خلال البحث عن الكلمات الحقيقية التي تتصاحب مع هذه الكلمة في كل مرة تظهر فيها داخل النص، مع اعتبار موقع هذه الكلمات والمشار إليه بالرمز d قبل الكلمة موضع البحث وبعدها، مع اعتبار أن $d(d=7)$، وهذه الكلمات الحقيقية التي يتم الحصول عليها يُنْظَر إليها على أنها الخصائص المنتَخَبة لهذه الكلمة. وهذه الخطوة يتم تنفيذها مرة واحدة.

(ب) الاختيار بين الخصائص: وفقًا لمجموعة الأكواد الدلالية للكلمة متعددة الدلالة موضع البحث يتم تحديد الكلمات التي تساعد في الحكم على معنى الكلمة متعددة الدلالة؛ وذلك وفقًا لقيمة درجة الحيرة (الإنتروبي) الذي يُعَبِّر عن خصائص الكلمات المُنْتَخَبة، وتمثل هذه الكلمات مجموعة الخصائص الدلالية التي تتصف بها هذه الكلمة. على سبيل المثال كلمة "材料" هي كلمة متعددة الدلالة لها ثلاثة أكواد دلالية (Ba06/Dk17/A103)، فتتم مرحلة اختيار الخصاص وعمليات القياس التي تتبعها وفقًا لهذه التصنيفات الدلالية الثلاثة. لذلك فإن هذه العملية هي عملية تفاعلية في إزالة اللبس.

(ج) وزن الخصائص: بعد اختيار الخصائص، يتم حساب قدرة كل خاصية على التوصيف الدلالي للكلمة موضع البحث، ونتيجة ذلك الحساب هي وزن كل خاصية من هذه الخصائص. ومن ذلك يتم تكوين مُتَّجِه للخصائص الدلالية لكل تصنيف دلالي تنتمي إليه الكلمة متعددة الدلالة.

شكل (٥-٥): نموذج إزالة اللبس الدلالي عن الكلمات

اعتمادًا على الفراغ الدلالي للكلمة داخل التركيب



(د) التعبير عن الخصائص: وفقًا لمجموعة الخصائص التي يتم اختيارها فيما سبق، يتم استخدام مُتَّجَه خصائص واحد للتعبير عن السياق الذي تتواجد فيه الكلمة متعددة الدلالة موضع البحث.

(هـ) حساب درجة التشابه: حيث تُحسب درجة التشابه بين مُتَّجَه الخصائص للكلمة موضع الدراسة والذي تم تحديده في الخطوة السابقة وبين مُتَّجه الخصائص لجميع التصنيفات الدلالية التي تنتمي إليها تلك الكلمة.

(و) تحديد معنى الكلمة: حيث يتم تحديد المعنى الحالي للكلمة من خلال التصنيف الدلالي الذي يحصل على أعلى درجة تشابه. فإذا لم يكن هناك مثل هذا التصنيف الدلالي، يتم التصعيد إلى الفئة الدلالية التي تنتمي إليها الكلمة، وإعادة عمليات إزالة اللبس إلى أن يتم التوصل إلى تحديد معنى الكلمة.

٤- نتيجة تجربة إزالة اللبس الدلالي عن الكلمات

قامت لي جوان تزي بتنفيذ هذه التجربة على ذخيرة نصوص إخبارية من (جريدة الشعب اليومية) وصل حجمها إلى ٧٢ ميجا بايت. وفي أثناء التجربة تم الاستعانة بمادة لغوية ثنائية التصنيف. وكان الهدف من الاختبار هو:

أ- التأكد من فاعلية الطريقة الحاسوبية سالفة الذكر في إزالة اللبس.

ب- استطلاع درجة إفادة منهجية الفراغ الدلالي للكلمة داخل السياق في تمييز دلالة الكلمات. وقد استخدَمَت لي تزى جوان الأسلوب التجريبي في اختبار كل من اللبس الكاذب واللبس الحقيقي للكلمات.

(١) اختبار اللبس الكاذب في معنى الكلمة

ما يُطلَق عليه "اللبس الكاذب" هو "الكلمة متعددة المعنى" التي تنشأ معانيها من انتماء كلمة أحادية الدلالة إلى كَودَيْن أو أكثر من أكواد التصنيف الدلالي. وباتباع طريقة إزالة اللبس التي نعرفها حاليًا، فإن من الممكن اعتبار تلك الكلمات أحادية الدلالة كلمات ذات تصنيف نحوي واحد، مثل: "收购"، و"修改" من الممكن أن يُكَوِّنا معًا زوجًا من الكلمات ذات اللبس الكاذب "修改/收购"، ويكون تصنيف اللبس الخاص بهما كما يلي **He03/Hg18**.

إن اختبار اللبس الكاذب هو أحد أنواع التجارب الحاسوبية التي تستخدم في إزالة اللبس (**Schutze 1992, Gale et al. 1992**)، وهذا الاختبار من شأنه أن يُقَلِّل الجهد الذي يستغرق في تجارب الترميز الدلالي للكلمات.

ويُجرَى بالتفصيل من خلال البدء بالبحث عن عدد المرات التي تظهر فيها كل كلمة من الكلمات التي تنتمي إلى اللبس الكاذب داخل ذخيرة الاختبار، ثم تُستَبْدَل تلك الكلمات بأزواج كلمات اللبس الكاذب. وهكذا فبعد إزالة اللبس الدلالي، يمكن استخدام ذخيرة الاختبار الأولى في حساب معدل الدقة في إزالة اللبس الدلالي. وبصورة عامة فإن استخدام هذه الطريقة التجريبية من شأنه أن يختبر مدى فاعلية أي طريقة في إزالة اللبس الدلالي.

وينقسم اختبار اللبس الكاذب إلى نوعين: اختبار مغلق، واختبار مفتوح. فالمادة اللغوية التي تُستَخدم في الذخيرة المُغلَقة يتم اختيارها من ذخيرة تجريبية؛ حيث يتم اختيار ٢٠٠ مثال لكل زوج من أزواج اللبس الكاذب بشكل عشوائي من الذخيرة. أما المادة المُستَخدَمة في الاختبار المفتوح فيتم اختيارها من ذخيرة من التصنيف نفسه، ولكنها خارج نطاق الذخيرة التجريبية، ويتم اختيار ١٠٠ مثال لكل زوج من أزواج اللبس الكاذب بشكل عشوائي. ويتم حساب معدل الدقة في إزالة اللبس الدلالي من المعادلة التالية:

معدل الدقة=	عدد الكلمات التي تم تحديد معناها بشكل صحيح
	العدد الإجمالي للكلمات التي تحمل لبسًا دلاليًا كاذبًا داخل ذخيرة الاختبار

ويَعْرِض الجدول ٥-١٧ عدد خمس نتائج اختبار للبس الدلالي الكاذب. وهذا لإثبات أن مُتَّجَهات التصنيف الدلالي التي تتكون من كلمات اللبس الدلالي الكاذب تتمتع بصفة الشمول، أما الجدول ٥-١٨ فيعرِض عدد المرات التي

تظهر فيها كل كلمة من أزواج اللبس الدلالي الكاذب والفئات الصغرى التي تنتمي إليها داخل ذخيرة التجريب. وقد أظهرت نتيجة التجربة أن:

أ- بالنسبة إلى متوسط معدل الدقة في إزالة اللبس الكاذب عن الكلمات، وصلت النسبة في كل من الاختبار المُغْلق والاختبار المفتوح إلى ٩٣,٥%، و٩٢,٦% على التوالي. وهذا يدل على فاعلية نموذج لي تزي جوان لإزالة اللبس الدلالي اعتمادًا على الفراغ الدلالي للسياق المحيط بالكلمة، والطريقة التفاعلية التي استخدمتها في اختيار الخصائص كانت مناسبة.

بِ- هناك بعض أزواج اللبس الكاذب، وعلى الرغم من أن عدد مرات ظهورها في ذخيرة التجريب كان ضعيفًا، فإن التصنيفات الدلالية الأصغر التي تنتمي إليها هذه الكلمات كانت تظهر بكثرة في ذخيرة التجريب، ومن ثَمَّ حصلت هذه الكلمات على نسبة دقة عالية في إزالة اللبس. ويبدو من ذلك أن البيانات الإحصائية التي يتم الحصول عليها باستخدام الكلمات أحادية الدلالة داخل الذخيرة، من الممكن أن تعكس بشكل أساسي الظروف العامة لتوزيع التصنيفات الدلالية داخل الذخيرة.

(٢) اختبار إزالة اللبس الحقيقي

اللبس الحقيقي يشير إلى الكلمات التي لها تعدد دلالي حقيقي؛ حيث يكون من المهم اختيار جزء من تلك الكلمات متعددة الدلالة من بين

التصنيفات المختلفة للكلمات لإجراء التجربة. ونظرًا إلى أن هـذه الكلمـات متعددة الدلالة لم يسبق التفكير فيها عند تكوين مُتَّجهات التـصنيف الـدلالي، لذلك لا تتضمن تجربة إزالة اللبس الحقيقي اختبارًا مُغْلَقًـا. وتبقـى معادلـة حساب معدل الدقة في إزالة اللبس الدلالي كما أوضحنا سابقًا.

جدول (٥-١٧): نتيجة تجربة إزالة اللبس الكاذب

أزواج كلمات اللبس الكاذب	معدل الدقة في الاختبار المغلق	معدل الدقة في الاختبار المفتوح
权利/事故	97.5%	94.0%
草案/责任	91.0%	89.5%
预算/预赛	98.0%	95.0%
收购/修改	93.0%	93.0%
颁发/参与	92.5%	87.0%
متوسط معدل الدقة	93.5%	92.6%

جدول (٥-١٨): إحصائية الكلمات التي تحمل لبسًا كاذبًا

أزواج كلمات اللبس الكاذب	تصنيف اللبس	عدد مرات الظهور في الذخيرة	عدد مرات ظهور التصنيف الدلالي الأصغر
权利/事故	Di21/Da01	959/979	2187/5088
草案/责任	Dk17/Di22	929/1563	4010/4177
预算/预赛	Hj29/hh07	176/841	4450/7914
收购/修改	He03/Hg18	788/954	1135/2383
颁发/参与	Hc11/Hi23	825/449	1472/3062

ونتيجة اختبار اللبس الدلالي الحقيقي يُظهِرها الجـدول ٥-١٩؛ حيـث تُشِير نتيجة التجربة أن فكرة استخدام تكوين الفراغ الدلالي للكلمات أحاديـة الدلالة وفقًا للتركيب الذي ترد فيه مع الكلمات الحقيقية داخل ذخيرة التجريب

منطقية. ومقارنةً بمنهجية إزالة اللبس دون الاعتماد على السياق نجد أن النموذج اللغوي الذي استخدمته لي تزي جوان يتمتع بمعدل دقة أعلى. وأهم ما في الأمر أن هذا النموذج اللغوي يتمتع بالقدرة على إجراء عمليات إزالة اللبس للكلمات على نطاق واسع، بالإضافة إلى أن هذه الطريقة في وصف دلالة الكلمات قد استُخْدِمَت من حيث المبادئ في لغات طبيعية أخرى غير اللغة الصينية.

جدول (٥-١٩): نتيجة التجربة المفتوحة في إزالة اللبس الحقيقي عن الكلمات

الكلمة متعددة الدلالة	تصنيف اللبس	عدد التجارب	عدد مرات ظهور التصنيف الدلالي الأصغر داخل الذخيرة	معدل الدقة في إزالة اللبس
材料	Dk17/Ba06/A103	791	422/1021/1913	81.7%
改	Ih02/Hg18/Hj66	2841	309/1135/1315	70.6%
表现	Jd06/Di20/Hj59	754	20/1500/1323	68.9%
发表	Hc11/Hi14/Jd03	2973	214/2943/5761	73.4%
建康	Ed43/Eb37	902	101/1056	70.1%
متوسط معدل الدقة				72.9%

٥- الخلاصة

(أ) إن نموذج إزالة اللبس الدلالي اعتمادًا على الفراغ الدلالي للكلمة داخل التركيب من الممكن أن يُقَلِّل أعباء العمل في الترميز الدلالي للكلمات أو بناء بنك معلومات ذي نطاق واسع حول إزالة اللبس الدلالي عن الكلمات.

(ب) بالنسبة إلى الفراغ الدلالي للكلمة داخل التركيب، فإنه كلما انخفضت قيمة مُتَّجه التصنيف الدلالي، عبر ذلك بصورة واقعية عن أحوال توزيع التصاحب للكلمات المترادفة داخل ذلك التصنيف الدلالي؛ ومن ثَمَّ يرتفع معدل الدقة في إزالة اللبس الدلالي. ومن المقترح أن تقتصر مستويات إزالة اللبس بهذه الطريقة على المستويين الثالث والرابع (أي التصنيف الأصغر والأصغر الفرعي بمعجم (غابة المترادفات).

(ج) إن جودة نتيجة إزالة اللبس الدلالي أو سوءها تتوقف على الخصائص النحوية للكلمات متعددة الدلالة. فبصورة عامة نجد أن نتيجة إزالة اللبس الدلالي عن الأسماء أفضل منها بالنسبة إلى الأفعال والصفات. وإذا تحدثنا عن الأفعال نجد أن نتيجة إزالة اللبس عن الأفعال متعددة الدلالة التي تقترن بمفعول به بسيط أفضل منها بالنسبة إلى الأفعال التي تقترن بمفعول به مركب (مثل المفعول به الجملة، والمفعول به المحوري الذي يشغل وظيفة المسند إليه بجانب كونه مفعولاً به).

(د) كما أن نتيجة إزالة اللبس الدلالي تتوقف على تصنيف اللبس بالنسبة إلى الكلمة متعددة الدلالة، فكلما قلت المسافة بين الفئات الدلالية التي تتبعها الكلمات متعددة الدلالة، ضعفت نتيجة إزالة اللبس.

إن بناء الفراغ الدلالي للكلمات انطلاقًا من السياق المصاحب لها اعتمادًا على ذخيرة كبيرة الحجم، يتناسب من حيث المبدأ مع معالجة أية كلمة حقيقية متعددة الدلالة، هذا بالإضافة إلى إمكانية تطبيقه مع أي لغة أخرى غير اللغة الصينية.

مـرفقـات

قائمة الرموز المستخدمة في التصنيف النحوي للكلمات

اسم N

اسم عام NG

اسم علم شخصي NF

اسم علم جغرافي NL

اسم مؤسسة NU

ظرف زمان T

ظرف مكان S

كلمة موضع F

فعل V

فعل مساعد VA

فعل رابط IV

فعل معبر عن الاتجاه VQ

فعل الكينونة VY

فعل الملكية VH

فعل مصدري VN

فعل مصدري متعدٍ VNN

فعل مصدري لازم VNO

أفعال عامة VG

فعل ذو مفعول مفرد VGN

فعل ذو مفعول مركب VGV

فعل لازم VGO

صفة A

كلمة حالية Z

تمييز B

عدد M

عدد أساسي MJ

عدد ترتيبي MX

أعداد أخرى MG

كلمة كمية Q

كلمة كمية للاسم MQ

كلمة كمية للفعل QV

ضمير R

حرف جر P

ظرف D

حرف عطف C

حرف عطف مُقَدَّم CF

حرف عطف أوْسَط CM

حرف عطف مُؤَخَّر CN

كلمات مساعدة U

الكلمة المساعدة "的" USDE

الكلمة المساعدة "地" USDI

الكلمة المساعدة "得" USDF

الكلمة المساعدة "似的" USSI

الكلمة المساعدة "所" USSU

الكلمة المساعدة "之" USZH

كلمة مساعدة مُعَبِّرة عن الزمن UT

كلمات مساعدة أخرى UX

كلمات اللهجة Y

الكلمات المعبرة عن الصوت O

كلمة تعجب E

سابقة H

لاحقة K

مَثَل I

اختصار J

تعبيرات شائعة L

أخرى X

سلاسل الرموز غير الرموز الصينية XCH

علامات الترقيم (كل منها يمثل تصنيفاً مستقلا)

فهرس المصطلحات الواردة في الكتاب

عربي – صيني

转移概率 - transition probability	احتمال التحول
出现概率	احتمال الظهور
词频统计	إحصاء معدل تكرار الكلمات
Z 分值 - Z-score	اختبار زد
测试 x^2	اختبار ك^2
切分歧义	تمييز اللبس
词义排歧 - Word Sense Disambiguation (WSD)	إزالة اللبس الدلالي
无指导的词义派歧	إزالة اللبس الدلالي غير الموجه
调查词义分布	استقراء التوزيع الدلالي للكلمة
体裁 - writing style	أسلوب الكتابة
语言习得 - language acquisition	اكتساب اللغة
定量统计 - quantative statistics	الإحصاء الكمي
封闭测试	الاختبار المغلق
开放测试	الاختبار المفتوح
统计测试	الاختبارات الإحصائية
语言使用 - language performance	الأداء اللغوي
索引 - search	الاستعلام
逐词索引 - concordance	الاستعلام الإحصائي السياقي عن الكلمات
词语检索 - word search	الاستعلام بالكلمة
文本检索 - text search	الاستعلام بالنص

关键词居中索引 - key word in context search	الاستعلام عن كلمة مفتاحية داخل السياق
疑问	الاستفهام
调查	الاستقراء
说明文体	الأسلوب التفسيري
议论文体	الأسلوب الجدلي
记叙文体	الأسلوب السردي
描写文体	الأسلوب الوصفي
系动词 - copula	الأسماء الموصولة
特殊化动词类	الأفعال التي لها سمات خاصة
情态动词	الأفعال المتصرفة
标准偏差 - standard deviation	الانحراف المعياري
词类	الأنواع النحوية للمفردات
数学分析	التحليل الرياضي
计量分析 - quantitive analysis	التحليل الكمي
计量语言分析 - quantative language analysis	التحليل الكمي للغة
多维分析 - multi-dimensional analysis	التحليل متعدد الاتجاهات
词法语法关联 - lexico-grammatical associations	التداخل المعجمي النحوي
简约性和不爱用的结构	التراكيب المختصرة والتي يَنْدُر استخدامها
索引行的排序	الترتيب المُفهْرَس للأسطر
机器翻译 - machine translation	الترجمة الآلية
标注 - tagging	الترميز
搭配 - collocation	التصاحب
左搭配词	التصاحب الأيسر
右搭配词	التصاحب الأيمن

自由组合 - free combination	التصاحب الحر
粘着组合 - bound combination	التصاحب اللصقي
搭配词 - collocations	التصاحبات
常用的搭配	التصاحبات شائعة الاستخدام
模糊范畴 - fuzzy category	التصنيف الضبابي
语义冲突	التضارب الدلالي
基本名词短语	التعبيرات الاسمية الأساسية
副词短语	التعبيرات الظرفية
形容词短语	التعبيرات الوصفية
计数 - counting	التعداد
聚类	التّعنقد
语义亲近 - affinity	التقارب الدلالي
自动识别	التمييز الآلي
词组边界自动识别	التمييز الآلي لحدود التعبيرات اللغوية
语料库自动分词	التمييز الآلي لكلمات الذخيرة
光电字符识别 - optical character recognition	التمييز الكهروضوني للحروف
同现 - co-occurrence	التوارد
词义分布	التوزيع الدلالي للكلمة
自然语言生成 - automatic natural language generation	التوليد الآلي للغات الطبيعية
逐词索引表 - key word in context (KWIC)	الجدول السياقي للكلمات
主谓谓语句	الجملة الإسنادية
状语	الحال
处所和时间状语	الحال المُعبّر عن المكان والزمان
多项状语	الحال المُوَسَّع

信息量 - entropy	الجمل المعلوماتي - الإنتروبي - مقياس درجة الفوضى
智能计算机	الحواسيب الذكية
话语 - discourse	الخطاب
同质的语料库 - homogeneous corpus	الذخيرة المتجانسة
异质的语料库 - heterogeneos corpus	الذخيرة المختلطة
系统的语料库 - systematic corpus	الذخيرة المنظومية
最大交集字段	السلسلة ذات الطول الأكبر في التداخل
上下文 - contexts	السياقات
语式	الشكل اللغوي
名词形式	الصيغ الاسمية
信息熵	الضجيج المعلوماتي
代名词	الضمير المُعَبِّر عن الاسم
代动词	الضمير المُعَبِّر عن الفعل
搭配尖峰	الطفرة التصاحبية
并列	العطف
统计相关 - correlational	العلاقات التبادلية
交集因子	العناصر المتداخلة
样本 - samples	العينات
词义空间	الفراغ الدلالي للكلمة
输出无关假设 - output-independence assumption	الفرض المستقل عن المخرجات
句子成分自动切分	الفصل الآلي لعناصر الجملة
检索 -- concordance	الفهرسة
语言能力 - language competence	القدرة اللغوية
阈值	القيمة المبدئية
短语词 - phrase word	الكلمات التعبيرية

搭配词	الكلمات المتصاحبة
关键词居中索引 - KWIC key words in context	الكلمات المفتاحية داخل السياق
高频词	الكلمات ذات معدل التكرار الأعلى
一词多义 - polysemy	الكلمات متعددة الدلالة
中心词	الكلمة المركزية
歧义	اللبس
真歧义	اللبس الحقيقي
伪歧义	اللبس الكاذب
通配符	اللواحق النحوية
被动	المبنى للمجهول
向量 - vectors	المُتَّجَهات
输出概率密度函数 - Output Probability Density Function	المتوالية المُعَبِّرَة عن كثافة مخرجات الاحتمال
定语	المُحَدِّدَات
条目	المداخل اللغوية
代表性 - representation	المرجعية
书面语	المستوى التحريري للغة
口语	المستوى الشفهي من اللغة
语域 - register	المستويات اللغوية
谓语	المُسْنَد
主语	المُسْنَد إليه
上下文信息	المعلومات السياقية
语境中的意义	المعنى السياقي
词汇特殊性	المفردات الخاصة
情态补语	المُكَمِّل المبين للحالة الشعورية
程度补语	المُكَمِّل المبين للدرجة

经验主义 - empiricism	المنهج التجريبي
理性主义 - rationalism	المنهج العقلي
转换生成语法理论 - transformational generative grammar	النحو التحويلي التوليدي
经验主义 - empiricism	النزعة التجريبية
理性主义 - rationalism	النزعة العقلية
比例	النسبة
段次比	النسبة العددية
段型比	النسبة النوعية
定语	النعت
多项定语	النعت المُوَسَّع
谓语形容词	النعوت الخبرية
定语形容词 - attributive adjectives	النعوت الوصفية
否定	النفي
语言模型 - language modeling	النمذجة اللغوية
词项 - lexical item	الوحدات المُعجمية
句法组成模板	أنماط التركيب النحوي
句型 - sentence pattern	أنماط الجمل
逐词索引软件	برمجيات الاستعلام عن الكلمات
检索系统	برمجية الإستعلام
索引程序 - Concordance program	..جيه فهرسة
引文条 - citation slips	بطاقات استشهاد ورقية
建立语料库 - corpus building	بناء الذخيرة
编纂 - compile	تجميع
自动句法分析	تحليل التركيب النحوي للجمل آليًا
话语分析 - discourse analysis	تحليل الخطاب
因素分析 - factor analysis	تحليل العوامل

文本至语音转换 -- text to sound conversion	تحويل النصوص المكتوبة إلى مكافئها الصوتي
存储语料 - storage of language materials	تخزين المادة اللغوية
降序排列	ترتيب تنازلي
标注 - tagging	ترميز
语料库标注 - corpus tagging	ترميز الذخائر اللغوية
词类标注	ترميز الكلمات من حيث التصنيف النحوي
词义标注 - Word sense tagging	ترميز معاني الكلمات
搭配	تصاحب
语料库设计 - corpus design	تصميم الذخيرة اللغوية
非基本名词短语	تعبيرات اسمية غير أساسية
介词短语	تعبيرات الجار والمجرور
第二语言的教学	تعليم اللغات للناطقين بغيرها
特殊用途的语言教学	تعليم اللغة للأغراض الخاصة
切分	تقطيع - تمييز
句子切分	تقطيع على مستوى الجمل
频次	تكرار
对文本实行电子编码	تكويد النصوص إلكترونيًا
语音识别	تمييز الأصوات اللغوية
分词	تمييز حدود الكلمات
语域变体 - register variation	تنوع المستوى اللغوي
索引表	جدول مفهرس
语料库的保护 - corpus protection	حماية الذخيرة اللغوية
主从关系特征	خصائص الإضافة
词语搭配的研究	دراسة التصاحبات اللغوية بين الكلمات

离散度	درجة التشتت
自由度	درجةالحُريَّة
维分 - dimention score	درجة المؤشر
搭配得离散度	درجة انتشار التصاحب
偏向	درجة تُحيُّز
数据噪音	درجة ضجيج البيانات
搭配的明显性	درجة وضوح علاقة التصاحب
凸显词	الكلمات ذات البروز الدلالي
动态语料库 - dynamic corpus	ذخائر لغوية ديناميكية متغيرة
规模较大的语料库	ذخائر لغوية فائقة الحجم
监督语料库 - monitor corpus	ذخائر لغوية لمراقبة التغيرات اللغوية
专用的语料库 - specialized corpus	ذخيرة الاستخدام المتخصص
训练语料库	ذخيرة تعلُّم
语料库 - corpus	ذخيرة لغوية
生语料库 - raw corpus	ذخيرة لغوية خام
计算机上的语料库 - computerized corpus	ذخيرة لغوية مُمَيْكنة
文本等级	رتبة النص
宏结构	سلاسل ذات تركيب كبير
交集型歧义切分字段	سلسلة كلامية تحتاج إلى إزالة اللبس من خلال فصل تداخل الحدود
交集字段	سلسلة متداخلة
状态形式	صيغة الحالة
交集字段的链长	طول سلسلة التداخل
耦合长度	طول محور التداخل
出现	ظهور

频次	عدد مرات التكرار
词频 - word frequency	عدد مرات تكرار الكلمة
语义无关	عدم الارتباط الدلالي
标点符号 - punctuation	علامات الترقيم
时、体标志	علامة الزمن والحالة
语义学 - semantics	علم الدلالة
语料库语言学 - corpus linguistics	علم الذخائر اللغوية
词法 - morphology	علم الصرف
历史比较语言学 - historical-comparative linguistics	علم اللغة التاريخي المقارن
语用学- pragmatics	علم اللغة التداولي - علم الاستخدام اللغوي
计算语言学 - computational linguistics	علم اللغة الحاسوبي
理论语言学 - theoretical linguistics	علم اللغة النظري
方言学 - dialectology	علم اللهجات
词典学 - lexicography	علم المعجمية
应用词汇学 - applied lexicology	علم المفردات التطبيقي
语法学 - grammar	علم النحو
分类心理学	علم النفس التصنيفي
词典编纂 - lexicography	علم تأليف المعاجم
样本 - sample	عينة
多为实值向量空间	فراغ المُتَّجَه الحقيقي متعدد الأبعاد
转换规则空间	فراغ قواعد التحويل
马尔科夫假设 - Markov assumption	فرض ماركوف
不及物动词 - transitive verb	فعل لازم
及物动词 - untransitiveverbs	فعل متعدي
助动词 - helpin verbs	فعل مساعد

名动词	فعل مصدري
不可带宾语的动词	فعل مصدري لازم
可带宾语的动词	فعل مصدري متعدي
趋向补语	فعل معبر عن الاتجاه
索引 - index	فهرس
数据库 - database	قاعدة بيانات
上下文有关规则	قواعد التحويل السياقية
相关程度 - strength of the relationship	قوة الارتباط
搭配强度	قوة التصاحب
阈值	قيمة الفرض المبدئي
负值	قيمة سالبة
语气词	كلمات اللهجة
结构助词	كلمات مساعدة
叹词	كلمة تعجب
状态词	كلمة حالية
量词	كلمة كمية
名量词	كلمة كمية للاسم
动量词	كلمة كمية للفعل
时态助词	كلمة مساعدة مُعَبِّرة عن الزمن
关键词居中 - key word in context	كلمة مفتاحية داخل السياق
方位词	كلمة موضع
互信息 - mutual information	كمية المعلومات المتبادلة
后缀	لاحقة
真歧义	لَبس حقيقي
伪歧义	لَبس زائف
复句	الجملة المركبة
多维向量	مُتَّجَه متعدد الأبعاد

评价函数	مُتواليات التقييم
函数	متوالية
限定性定语	مُحَدَّدات للتقييد
区别性定语	مُحَدَّدات للتمييز
描写性定语	مُحَدَّدات للوصف
耦合段	محور التداخل
词条 - entry	مَدخل
公式	معادلة
语料库的加工 - corpus processing	معالجة الذخيرة
然语言处理 - natural language processing	معالجة اللغات الطبيعية
召回率	معدل الارتداد
频率 - frequency	معدل التكرار
静态频率	معدل التكرار الثابت
真正频率 - True frequency	معدل التكرار الحقيقي
动态频率 - variant frequency	معدل التكرار المتغير
期望频率 - xpected frequency	معدل التكرار المتوقع
准确率	معدل الدقة
词形在语料库中的出现频率	معدل ظهور الكلمةداخلالذخيرة
补语	مكمل المعنى
数据文件 - data file	ملف البيانات
逐词索引文件 - concordance file	ملف سياقات الكلمات
维..................	مؤشر
窗口 - N-gram	نافذة
观察窗口	نافذة الاستقراء
想象性文本 - Imaginative text	نصوص إبداعية
知识性文本 - Informative text	نصوص معلوماتية

形式语法 - Formal Grammar	نظرية النحو الصوري
机器翻译系统 - mahcine translation systems	نُظم الترجمة الآلية
转换规则模板	نماذج قواعد التحويل
N 元模型	نموذج العنصرN
隐马尔可夫模型 - Hidden Markov Model (HMM)	نموذج ماركوف الكامن
语言工程 - language engineering	هندسة اللغة
平衡 - balance	التوازن
均值	المتوسط الحسابي
初始标注模块	وحدة الترميز المبدئي
权值 - weight	وزن
微结构	سلاسل ذات تركيب دقيق
文本媒介	وسيط النشر
召回率	معدل التغطية

فهرس المصطلحات الواردة في الكتاب

صيني- عربي

被动	المبنى للمجهول
比例	النسبة
编纂	compile - تجميع
标点符号	punctuation - علامات الترقيم
标注	tagging - الترميز
标准偏差	standard deviation - الانحراف المعياري
并列	العطف
补语	مكمل المعنى
不及物动词	transitive verbs - فعل لازم
不可带宾语的动词	فعل مصدري لازم
常用的搭配	التصاحبات شائعة الاستخدام
程度补语	المُكمّل المبين للدرجة
出现	ظهور
出现概率	احتمال الظهور
初始标注模块	وحدة الترميز المبدئي
处所和时间状语	الحال المُعبّر عن المكان والزمان
窗口	N-gram - نافذة
词典编纂	lexicography - علم تأليف المعاجم
词典学	lexicography - علم المعجمية
词法	morphology - علم الصرف
词法语法关联	التداخل المعجمي النحوي - lexico-grammatical associations
词汇特殊性	المفردات الخاصة
词类	الأنواع النحوية للمفردات

词类标注	ترميز الكلمات من حيث التصنيف النحوي
词频	wordfrequency - عدد مرات تكرار الكلمة
词频统计	إحصاء معدل تكرار الكلمات
词条	entry - مَدْخَل
词项	lexical item - الوحدات المُعجمية
词形在语料库中的出现频率	معدل ظهور الكلمة داخل الذخيرة
词义标注	Word - ترميز معاني الكلمات sensetagging
词义分布	التوزيع الدلالي للكلمة
词义空间	الفراغ الدلالي للكلمة
词义排歧	Word) - إزالة اللبس الدلالي SenseDisambiguation (WSD
词语搭配的研究	دراسة التصاحبات اللغوية بين الكلمات
词语检索	word search - الاستعلام بالكلمة
词组边界自动识别	التمييز الألي لحدود التعبيرات اللغوية
存储语料	storage - تخزين المادة اللغوية oflanguage materials
搭配	collocation - التصاحب
搭配	تصاحب
搭配词	collocations - التصاحبات
搭配词	الكلمات المتصاحبة
搭配得离散度	درجة انتشار التصاحب
搭配的明显性	درجة وضوح علاقة التصاحب
搭配尖峰	الطفرة التصاحبية
搭配强度	قوة التصاحب
代表性	representation - المرجعية
代动词	الضمير المُعبِّر عن الفعل
代名词	الضمير المُعبِّر عن الاسم

第二语言的教学	تعليم اللغات للناطقين بغيرها
定量统计	quantativestatistics - الإحصاء الكمي
定语	المُحَدِّدَات
定语	النعت
定语形容词	attributiveadjectives - النعوت الوصفية
动量词	كلمة كمية للفعل
动态频率	variantfrequency - معدل التكرار المتغير
动态语料库	dynamic - ذخائر لغوية ديناميكية متغيرة corpus
短语词	phrase word - الكلمات التعبيرية
段次比	النسبة العددية
段型比	النسبة النوعية
对文本实行电子编码	تكويد النصوص إلكترونيًا
多为实值向量空间	فراغ المُتَّجَه الحقيقي متعدد الأبعاد
多维分析	multi- - التحليل متعدد الاتجاهات dimensionalanalysis
多维向量	مُتَّجَه متعدد الأبعاد
多项定语	النعت المُوَسَّع
多项状语	الحال المُوَسَّع
方位词	كلمة موضع
方言学	dialectology - علم اللهجات
非基本名词短语	تعبيرات اسمية غير أساسية
分词	تمييز حدود الكلمات ،
分类心理学	علم النفس التصنيفي
封闭测试	الاختبار المغلق
否定	النفي
负值	قيمة سالبة
复句	الجملة المركبة

副词短语	التعبيرات الظرفية
高频词	الكلمات ذات معدل التكرار الأعلى
公式	معادلة
关键词居中	كلمة مفتاحية داخل السياق - key wordin context
关键词居中索引	الاستعلام عن كلمة مفتاحية داخل السياق - key word in context search
观察窗口	نافذة الاستقراء
光电字符识别	التمييز الكهروضوني للحروف - opticalcharacter recognition
规模较大的语料库	ذخائر لغوية فائقة الحجم
函数	متوالية
宏结构	سلاسل ذات تركيب كبير
后缀	لاحقة
互信息	كمية المعلومات المتبادلة - mutualinformation
话语	الخطاب - discourse
话语分析	تحليل الخطاب - discourse analysis
机器翻译	الترجمة الألية - machine translation
机器翻译系统	نُظم الترجمة الألية - mahcinetranslation systems
基本名词短语	التعبيرات الاسمية الأساسية
及物动词	فعل متعدي - untransitiveverb
计量分析	التحليل الكمي - quantitive analysis
计量语言分析	التحليل الكمي للغة - quantativelanguage analysis
计数	التعداد - counting
计算机上的语料库	ذخيرة لغوية مُمَيْكَنة - computerized corpus

计算语言学	علم اللغة الحاسوبي - computationallinguistics
记叙文体	الأسلوب السردي
监督语料库	ذخائر لغوية لمراقبة التغيرات اللغوية - monitor corpus
检索	الفهرسة - concordance
检索系统	برمجية الاستعلام
简约性和不爱用的结构	التراكيب المختصرة والتي يندر استخدامها
建立语料库	بناء الذخيرة - corpus building
降序排列	ترتيب تنازلي
交集型歧义切分字段	سلسلة كلامية تحتاج إلى إزالة اللبس من خلال فصل تداخل الحدود
交集因子	العناصر المتداخلة
交集字段	سلسلة متداخلة
交集字段的链长	طول سلسلة التداخل
结构助词	كلمات مساعدة
介词短语	تعبيرات الجار والمجرور
经验主义	المنهج التجريبي - empiricism
经验主义	النزعة التجريبية - empiricism
静态频率	معدل التكرار الثابت
句法组成模板	أنماط التركيب النحوي
句型	أنماط الجمل - sentence pattern
句子成分自动切分	الفصل الآلي لعناصر الجملة
句子切分	تقطيع على مستوى الجمل
聚类	التعنقد
均值	المتوسط الحسابي
开放测试	الاختبار المفتوح
可带宾语的动词	فعل مصدري متعدي

口语	المستوى الشفهي من اللغة
离散度	درجة التشتت
理论语言学	theoretical - علم اللغة النظري linguistics
理性主义	rationalism - المنهج العقلي
理性主义	rationalism - النزعة العقلية
历史比较语言学	historical- علم اللغة التاريخي المقارن comparative linguistics
量词	كلمة كمية
N 元模型	نموذج العنصر N
马尔科夫假设	Markov assumption - فرض ماركوف
描写文体	الأسلوب الوصفي
描写性定语	مُحَدِّدات للوصف
名词形式	الصيغ الاسمية
名动词	فعل مصدري
名量词	كلمة كمية للاسم
模糊范畴	fuzzy category - التصنيف الضبابي
耦合段	محور التداخل
耦合长度	طول محور التداخل
偏向	درجة تَحَيُّز
频次	تكرار
频次	عدد مرات التكرار
频率	frequency - معدل التكرار
平衡	balance - التوازن
评价函数	مُتَواليات التقييم
期望频率	معدل التكرار المتوقع - expectedfrequency
歧义	اللبس

切分	تقطيع - تمييز
切分歧义	تمييز اللبس
情态补语	المُكمِّل المبين للحالة الشعورية
情态动词	الأفعال المتصرفة
区别性定语	مُحَدِّدات للتمييز
趋向补语	فعل معبر عن الاتجاه
权值	weight - وزن
然语言处理	naturallanguage - معالجة اللغات الطبيعية processing
上下文	contexts - السياقات
上下文信息	المعلومات السياقية
上下文有关规则	قواعد التحويل السياقية
生语料库	raw corpus - ذخيرة لغوية خام
体标志、时	علامة الزمن والحالة
时态助词	كلمة مساعدة مُعَبِّرة عن الزمن
书面语	المستوى التحريري للغة
输出概率密度函数	المتوالية المُعَبِّرَة عن كثافة مخرجات الاحتمال Output ProbabilityDensity - Function
输出无关假设	output- الفرض المستقل عن المخرجات independence assumption
数据库	database - قاعدة بيانات
数据文件	data file - ملف البيانات
数据噪音	درجة ضجيج البيانات
数学分析	التحليل الرياضي
说明文体	الأسلوب التفسيري
索引	search - الاستعلام
索引	index - فهرس
索引表	جدول مفهرس

索引程序	Concordance program - برمجية فهرسة
索引行的排序	الترتيب المُفهرَس للأسطر
叹词	كلمة تعجب
特殊化动词类	الأفعال التي لها سمات خاصة
特殊用途的语言教学	تعليم اللغة للأغراض الخاصة
体裁	writing style - أسلوب الكتابة
条目	المداخل اللغوية
调查	الاستقراء
调查词义分布	استقراء التوزيع الدلالي للكلمة
通配符	اللواحق النحوية
同现	co-occurrence - التوارد
同质的语料库	homogeneouscorpus - الذخيرة المتجانسة
统计测试	الاختبارات الإحصائية
统计相关	correlational - العلاقات التبادلية
凸显词	الكلمات ذات البروز الدلالي
微结构	سلاسل ذات تركيب دقيق
维	مؤشر
维分	dimention score - درجة المؤشر
伪歧义	اللبس الكاذب
伪歧义	لبس زائف
谓语	المُسنَد
谓语形容词	النعوت الخبرية
文本等级	رتبة النص
文本检索	text search - الاستعلام بالنص
文本媒介	وسيط النشر
文本至语音转换	تحويل النصوص المكتوبة إلى مكافئها الصوتي text to sound conversion -
文学与语言计算	literary - الدراسات الحاسوبية للأدب واللغة

	and linguistic computing
无指导的词义派歧	إزالة اللبس الدلالي غير الموجه
系动词	copula - الأسماء الموصولة
系统的语料库	systematiccorpus - الذخيرة المنظومية
限定性定语	مُحَدِّدات للتقييد
相关程度	strength of the relationship - قوة الارتباط
想象性文本	Imaginative text - نصوص إبداعية
向量	vectors - المُتَّجَهات
信息量	الجِمَل المعلوماتي - الإنتروبي - مقياس درجة الفوضى - entropy
信息熵	الضجيج المعلوماتي
形容词短语	التعبيرات الوصفية
形式语法	Formal Grammar - نظرية النحو الصوري
训练语料库	ذخيرة تعلُّم
样本	samples - العينات
一词多义	polysemy - الكلمات متعددة الدلالة
疑问	الاستفهام
议论文体	الأسلوب الجدلي
因素分析	factor analysis - تحليل العوامل
引文条	citationslips - بطاقات استشهاد ورقية
隐马尔可夫模型	Hidden Markov Model (HMM) - نموذج ماركوف الكامن
应用词汇学	appliedlexicology - علم المفردات التطبيقي
右搭配词	التصاحب الأيمن
语法学	grammar - علم النحو
语境中的意义	المعنى السياقي

语料库	corpus - ذخيرة لغوية
语料库标注	corpustagging - ترميز الذخائر اللغوية
语料库的保护	corpus - حماية الذخيرة اللغوية protection
语料库的加工	corpus processing - معالجة الذخيرة
语料库设计	corpus design - تصميم الذخيرة اللغوية
语料库语言学	corpus linguistics - علم الذخائر اللغوية
语料库自动分词	التمييز الآلي لكلمات الذخيرة
语气词	كلمات اللهجة
语式	الشكل اللغوي
语言工程	language engineering - هندسة اللغة
语言模型	language modeling - النمذجة اللغوية
语言能力	language competence - القدرة اللغوية
语言使用	language performance - الأداء اللغوي
语言习得	language acquisition - اكتساب اللغة
语义冲突	التضارب الدلالي
语义亲近	affinity - التقارب الدلالي
语义无关	عدم الارتباط الدلالي
语义学	semantics - علم الدلالة
语音识别	تمييز الأصوات اللغوية
语用学	علم اللغة التداولي - علم الاستخدام اللغوي - pragmatics
语域	register - المستويات اللغوية
语域变体	تنوع المستوى اللغوي
阈值	القيمة المبدئية
阈值	قيمة الفرض المبدئي
Z 分值	Z-score - اختبار زد
粘着组合	bound combination - التصاحب اللصقي

召回率	معدل الارتداد
召回率	معدل التغطية
真歧义	اللبس الحقيقي
真歧义	لبْس حقيقي
真正频率	معدل التكرار الحقيقي - Truefrequency
知识性文本	نصوص معلوماتية - Informative text
智能计算机	الحواسيب الذكية
中心词	الكلمة المركزية
逐词索引	الاستعلام الإحصائي السياقي عن الكلمات - concordance
逐词索引表	الجدول السياقي للكلمات - key word in) context (KWIC
逐词索引软件	برمجيات الاستعلام عن الكلمات
逐词索引文件	ملف سياقات الكلمات - concordance file
主从关系特征	خصائص الإضافة
主谓谓语句	الجملة الإسنادية
主语	المُسْنَد إليه
助动词	فعل مساعد - helping verb
专用的语料库	ذخيرة الاستخدام المتخصص - specialized corpus
转换规则空间	فراغ قواعد التحويل
转换规则模板	نماذج قواعد التحويل
转换生成语法理论	النحو التحويلي التوليدي - transformational generative grammar
转移概率	احتمال التحول - transition probability
状态词	كلمة حالية
状态形式	صيغة الحالة
状语	الحال

准确率	معدل الدقة
自动句法分析	تحليل التركيب النحوي للجمل آليًا
自动识别	التمييز الآلي
自然语言生成	التوليد الآلي للغات الطبيعية - automatic language generation
自由度	درجةالحُريَّة
自由组合	التصاحب الحر - free combination
最大交集字段	السلسلة ذات الطول الأكبر في التداخل
左搭配词	التصاحب الأيسر

قائمة المراجع

[1]丁信善，《语料库语言学的发展及研究现状》，当代语言学，1998，1。

دينغ شين شان، "علم الذخائر اللغوية.. مراحل التطور وواقع البحوث الحالية"، مجلة ((علم اللغة المعاصر))، العدد الأول، ١٩٩٨.

[2] Mc Enery, T,Wilson. A,Corpus Linguistics, Edinburgh University Press, 1996.

[3] Crystal, D, Stylistic profiling, in Aijmer&Altenberg, 1991, pp. 221-238.

[4] Preyer, W, The Mind of a Child, New York: Appletion, 1889.

[5] Stern., Psychology of Early Childhood up to Six Years Old Age, New York: Holt, 1924.

[6] N. Francis, A tagged corpus-problems and prospects, in Green-baum et al. 1980, pp.192-209. New York: Longman.

[7] G. Kennedy, Preferred ways of putting things, in Svartvik, 1992, pp.335-373.

[8] C. Fries, The Structure of English: An Introduction to the Construction of Sentences, New York: Harcourt-Brace.

[9] Chomosky, N. Syntactic Structures, The Hague: Mouton,1957.

[10] Halliday, M. A. K. Corpus Studies and Probabilistic Grammar, in Aijmwe and Altenberg, 1991, pp.30- 43.

[11] J. Svartvik, Directions in Corpus Linguistics, Berlin: Mouton de Cruyter,1992.

[12] Halliday, M. A. K. Corpus Studies and Probabilistic Grammar, in Aijmer and Alterberg, 1991, pp.30-43.

[13] Govindankutty,A., The computer and Dravidian linguistics, ALLC bulletin, 1973.

[14] Johansson,S., Continuity and changes in the encoding of computer corpora, in Oostdijk&Hann, 1994,pp.13-31.

[15] Kennedy,G.,An Introduction to Corpus Linguistic, London&New York: Longdon,1998.

[16] Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985) AComprehesive Grammar of the English Language, Lodon: Longman.

[17] Sinclair, J. (1991) The automatic analysis of corpora,InSvartvik 1992, pp. 379-397.

[18] G. Leech, (1991) "The State of Art in Corpus Linguistics", in Aijmer and altenberg 1991, pp.8-29.

[19] Summers, D. Longman/Lancaster English Language Corpus: Criteria and Design, Harlow: Longman,1991.

[20]张普《关于大规模真实文本语料库的基点理论考察》，语言文字与应用，1999，1。

جانغ بو، "بعض الآراء التنظيرية حول الذخائر اللغوية كبيرة الحجم"، مجلة ((دراسات تطبيقية في اللغة))، العدد الأول،١٩٩٩.

[2١]Runddle, M., Stock. P., The corpus revolution, English Today, 1992.

[22]Biber, D. 1993, 'Representativness in Corpus Design', Literary and Linguistic Computing vol.8, No. 4:pp. 243-257.

[23] Sinclair, J. Corpus, Concordence, Collocatiion. OxfordUniversity, 1991.

[24]马少平《脱机手写体汉字识别方法与系统》，博士论文，清华大学计算机科学与技术系，1997，5。

ما شاو بينغ، ((طرق تمييز الرموز الصينية المكتوبة بخط اليدوأنظمتها))، رسالة دكتوراة، قسم علوم وتكنولوجيا الحاسب الآلي بجامعة تشين خوا، مايو، ١٩٩٧.

[25]国家语言文字工作委员会汉字处《现代汉语常用字表》，语文出版社。

قسم الرموز الصينية باللجنة القومية للأعمال الخاصة باللغة التحريرية، ((مسرد الرموز شائعة الاستخدام في اللغة الصينية الحديثة))، دار نشر اللغة والأدب.

[26] Walker, D.E., The Ecology Of Language, in The Proceedings of International Workshop on Electronic. Dictionaries,1990,OisoJapan.

[27] Quirk, R. ,Towards a description of English Usage, Transaction of Philological Society, pp.40-61.

[28] Francis, W. N. ,Kucera, H. Frequency Analysis of English Usage: Lexicon and Grammar Boston: Houghton Miffin.

[29] Greene, B. B. Rubin, G. M. Automatic grammatical tagging of English, Providence, R. I.: Department of Linguistics, BrownUniversity.

[30] Brown Corpus: http:// khnt. hit. uib. no/icame/ manuals/ brown/ index.htm.

[31] Johanson, S., Atwell, E., Carsid, R. And Leech, G. 1986. The tagged LOB corpus: User's Manual, Norwegian Computing Center for The Humanities, Bergen.

[32] Johanson, S. (ed.) Computer Corpora in English Language Research, Bergen: Norwegian Computing Centre for the Humanities, 1982.

[33] Johanson, S., Hofland, L.,FreauencyAnalyais of English Vocabulary and Grammar 2 vols. Oxford: Clarendon Press.

[34] Svartvik, J. (ed.) The London-Lund Corpus of Spoken English: Description and Research Lund: Lund Studies in English 82. LundUniversity Press.

[35] The Bank of English, http:// titania. Cobuild.Collins.co. uk/ boe- info- html.

[36]Collins COBUILD English Dictionary, Collins COBUILD.

[37]Della Summers, Longman Group UK, Longman/Lancaster English Language Corpus-Criteria and Design, International.Journal of Lexicography, Vol. 6 No. 3: pp. 181-208.

[38]British National Corpus, Http:// info.ox.ac.uk./Bnc.

[39] The International Corpus of English http://www.ucl.ac.uk/ English-usage/ice.htm.

[40] 陈鹤琴，《语体文应用字汇》，商务印书馆，1982。

تشن خه تشين، ((قائمة الرموز المستخدمة في الأساليب اللغوية المختلفة))، دار النشر التجارية، ١٩٨٢.

[41]

王还、常宝儒等《现代汉语频率词典》，北京语言学院出版社，1986。

وانغ خواي، وتشانغ باو رو، وآخرون، ((معجم معدل تكرار الكلمات في اللغة الصينية الحديثة))، دار نشر معهد اللغات الأجنبية ببكين، ١٩٨٦.

[42]陈原，《现代汉语定量分析》，上海教育出版社，1989。

تشن يوان، ((التحليل الكمي للغة الصينية الحديثة))، دار النشر التعليمية بشنغهاي، ١٩٨٢.

[43]台湾中央研究院平衡语料库，http://godel.iis.sinica.edu.tw

الذخيرة المتوازنة لمعهد الدراسات المركزي بتايوان،

http://godel.iis.sinica.edu.tw

[44] Huang, Chu-Ren and Keh-Jiann Chen, Modern and Classical Chinese Corpora at Academic Text Databases for Natural Language Proccessing and linguistic Computing, Presented at the Sixth CODATA Task Group Meeting on the Survey of Data Sources in Asian-Oceanic Countries (Taipei: Academic Sinica, 1994).

[45] Huang, Chu-Ren and Ken-Jiann Chen, Modem and Classical Chinese Corpora at Academic Sinica Text Databases for Natural Lamguage Processing and Computing, Presented at the Sixth CODATA Task group meeting on the Survey of Data Sources in Asian-Oceànic Countries (Taipei: Academic Sinica, 1994.

[46] K. J. Chen, C. R. Huang,L.P. Chang, H.L. Hsu, 1996, "SINICA CORPUS: Design Methodology for Balanced Corpora" Proceedings of PACLICLL, pp. 167-176, Seoul,Korea.

[47]香港城市大学语言咨询科学研究中心《中文各地区共时词语研究报告》，1998，5。

مركز الاستشارات اللغوية العلمية بجامعة المدينة بهونج كونج، ((تقرير بحثي تزامني عن المفردات في خمس مناطق ناطقة باللغة الصينية))، دار النشر التجارية، ١٩٨٢.

[48]孙茂松、黄建平等《现代汉语语料库系统鉴定会文件》，1996，1。

سوين ماو سونغ، وخوانغ جيان بينغ،وآخرون، ((محضر الاجتماع الخاص باعتماد برمجية ذخيرة اللغة الصينية الحديثة))،يناير، ١٩٩٦.

[49]王建新，《索引软件：语料库语言学的有利工具》，当代语言学，1998，1。

وانغ جيان شين،"برمجيات الفهرسة: أدوات مفيدة في علم الذخائر اللغوية"، مجلة ((علم اللغة المعاصر))، عدد يناير، ١٩٩٨.

[50]Biber, D. ,Finegan,E., Intra-textual variation within medical research articles, in Oosdijk& de Haan, pp.201-222

[51]Kjellmer, G. (1991) 'A mint of phrases', in aijmer and Altenberg 1991,pp. 111-127.

[52]朱雪龙、艾红梅《应用信息论基础》，清华大学内部教材，1 998。

جو شوي لونغ، وآي خونغ مي، ((مبادئ نظرية المعلومات التطبيقية))، مقرر دراسي خاص بجامعة تشين خوا، ١٩٩٨.

[53]Leech,G.(1992),Corpora ad theories of linguistic performance,inJ.Svartvik 1992,pp. 149-163.

[54]LariK.Young S. Applications of stochastic context-free grammars using the inside-outside algorithm,In: Computer Speech & Language 1991,5,pp.237-257.

[55] Pereira,F.,Schabes,Y. Inside-outside reestimation from partially bracket corpora,In: Proceedings of the 30th Annual Meeting of the Association for Computational Liguistics, University of Delaware, Newark, Delaware, USA, 1995, pp.128-135.

[56]EngenceCharniak, Statistical Language Learning, The MIT Press, London, England.

[57] John Sinclair, Longuageindepenent statistical software for corpus exploration. Coputers and the Hemanities, 1998, Vol.31, pp. 229-255.

[58]Garside,R., Leech,G., McEnery, A., Corpus Annotation, Longman & New York,1997.

[59]

刘开瑛《中文文本自动分词和标注》，北京：商务印书馆，2000。

ليو كاي يينغ، ((التمييز الآلي لحدود الكلمات داخل النصوص الصينية وترميزها))، بكين: درا النشر التجارية، ٢٠٠٠.

[60]Ellegard, A. (1978) The syntactic structure of English texts: A computer-based study of four kinds of text in Brown University Corpus. Guöteborg: Gothenburg Studies in English 43.

[61] Van Halteren, H., Oostdijk, N.,Towards a syntactic database: The TOSGA analysis system> In Arts et al> 1993, pp. 145-162.

[63] The Penn Treebank Project,

http://www.cis.upenn.edu/~treebank/home.html.

[64]Chelba, C. , Exploiting Syntactic Language Structural for Language Odeling, A Dissertation of JohnHopskinUniversity, January, 2000.

[65]Collins, M. J., A New Statstical Parser Based on Lexical Dependencies, The 34th Annual Meeting of the ACL, Santa Cruz, California 1996.

[66]Jelinek, F., Laffertyy, J. And Mercer, R., et al, Decision Tree Parsing Using a Hidden Derivation Model, the proceedings of the 1994 Human Language Technology Workshop, pp. 272-277.

[67] Richardson, S. D., Dolan, W. B., Vandewende, L., MindNet: Aquiring and Struturing Semantic Information from Text, ACL'98, vol. 2, pp.1098-1102.

[68] Church K., Gale, W., Hanks, P. And Hindle, D., Using statistic in lexical analysis, in Souter and Atwell 1993, pp. 85-96.

[69] Schmidt, K. M. Qualitative and quantitive research approaches to English constructions, in Souter and Atwell 1993, pp. 85-96.

[70]Mindt, D. Syntactic evidence for semantic distinctions in english, Aijmer and Altenberg 1991, pp. 182-196.

[71]Stenstorm, A. B. Carry on signals in English conversation, in Meijs 1987, pp. 87-119.

[72]赵淑华《现代汉语句型统计与研究》，成果报告，北京语言文化大学，1995，4，10。

جاو شو خوا، ((إحصاء أنماط الجمل في اللغة الصينية الحديثةودراستها))، تقرير عن نتائج بحثية، جامعة اللغات والثقافة ببكين، إبريل، ١٩٩٥.

[73] AHI 语料库：http://ref.umdl.umich.edu/a/ahd/sample.htm

ذخيرة إيه إتش آي AHI

http://ref.umdl.umich.edu/a/ahd/sample.htm

[74] Carroll, J.B., Davies, P. And Richman, b. The American heritage Word Frequency Book, New York: American Heritage Publishing Co.

[75] Collins Cobuild English Language Dictionary, 1987, London, Collins.

[76]黄居仁、陈克健等《国语日报量词典》，台北：国语日报社，1997。

خوانغ جو رن، وتشن كه جيان، ((معجم الكلمات الكمية المستخدمة في جرائد تايوان))، تاي بيه، دار نشر الجرائد التايوانية،١٩٩٧.

[77] Firth, J. R., A synopsis of linguistic theory, 1930-1955, in studies in Linguistic Analysis Oxford: Blachwell.

[78] Essex, (new ed) Longman dictinary of contemporary English, England: Longman.

[79] Summers, D. Longman language Activator, 1993, Longman.

[80] Biber, D., Finegan , E., On the exploitation of computerized corpora in variation studies, in Aijmer&Altenberg, 1991, pp. 204-220.

[81] Benson, M., Benson, E. And Ilson, R., The BBI Combinatory Dictionary of English, Amsterdam: John Benjamins Publishing Co., 1986.

[82] Benson, m., A Combinatory Dictionary Of English, Dictionaries:Journal Of The Dictionary Society Of North America, 7.

[83]
张寿康、林杏光《现代汉语实词搭配词典》，北京：商务印书馆，1992。

جانغ شو كانغ، ولين تشي جوانغ، ((معجم تصاحبات الكلمات الحقيقية في اللغة الصينية الحديثة))، بكين، دار النشر التجارية، ١٩٩٢.

[84]Choueka, Y., Klein, T., and Neuwitz, E., Automatic retrieval of frequent idiomatic and collocation expressions in a large corpus, Journal of Literary and linguistic Computing, 4.

[85] Church, K., Hanks, p., Word association, Mutual Information and Lexicography. In proceedings of 27th Annual meeting of Association for Computational Linguistics, 1989, pp. 76-83.

[86] Smadja, f., Retrieving Collocation from Text: Xtract, Computational linguistics, vol. 19, No.1.

[87]
孙茂松等《汉语搭配定量分析初探》，中国语文，1997，1
。

سوين ماو سونغ وآخرون، "دراسة أولية عن التحليل الكمي للتصاحبات اللغوية في اللغة الصينية"، مجلة ((اللغة والأدب في الصين))، يناير، ١٩٩٧.

[88] Ahrens, Kathleen and Chu-Renhuang, Classifiers and semantic Type Coercion: Motivating a new Classification of classifiers, In. B.-S. Park and J. B. Kim.eds. Proceeding Of the 11th pacific Asia Conference on Language, Information and Computation (Seoul: Kyung Hee University, 1996), pp. 1-10.

[89] Shannon, C. A mathematical theory of communications, Bell System Technical Journal, 1949.27, pp. 623-656.

[90] Bible, D., Conrad, S. And Reppen, R., Corpus-based Approaches to issues in applied linguistics, Applied Linguistics vol. 15, No.2, pp. 169-189.

[91]Biber, D., Variation across Speech and Writing, Cambridge: CambridgeUnivesity Press, 1988.

[92]

刘开瑛《自动分词与词性标注评测》，计算机世界，评测专版，1996，3，25。

ليو كاي يينغ، ((دراسة لتقييم منهجيات التقسيم الآلي لحدود الكلمات الصينية والترميز الآلي لأنواع الكلمات))، بكين، دار نشر عالم الكمبيوتر، نسخة خاصة للتقييم، مارس، ١٩٩٦.

[93]

孙茂松、黄昌宁等《零用汉字二元语法关系解决汉语自动分词中交集型歧义》，计算机研究与发展，1997年，第34卷第5期。

سوين ماو سونغ، وخوانغ تشانغ نينغ، وآخرون، "استخدام نحو العلاقات الثنائية بين الرموز الصينية في إزالة اللبس المركب عند التمييز الآلي لحدود الكلمات الصينية"، مجلة ((دراسات في تطوير علوم الحاسب))، العدد الخامس من المجلد ٣٤، ١٩٩٧.

[94]

吴芳芳《自动分词中歧义字段切分方法研究》，硕士论文，山西大学，1998。

وو فانغ فانغ، ((دراسات في منهجيات تقسيم مقاطع الرموز الصينية التي تمثل لبسًا في أثناء التمييز الآلي لحدود الكلمات الصينية))، رسالة ماجستير، جامعة شان شي، ١٩٩٨.

[95]

左正平《汉语自动分词中的若干问题》，清华大学计算机科学与技术系硕士论文，1998，6。

زُوَه جنغ بينغ، ((بعض مشكلات التمييز الآلي لحدود الكلمات الصينية))، رسالة ماجستير بقسم علوم الحاسب والتكنولوجيا بجامعة تشين خوا، يونيه، ١٩٩٨.

[96]

孙茂松、左正平《汉语真实文本中的交集型歧义》，汉语计量与计算研究，1998。

سوين ماو سونغ، وزُوَه جنغ بينغ، "لبس السلاسل المتداخلة في النصوص الصينية الحقيقية"، ((دراسات إحصائية وحاسوبية عن اللغة الصينية))، ١٩٩٨.

[97] Church, K., A stochastic parts program and noun phrase parser for unrestricted text, In: Proceedings of the Second Conference on applied Natural Language Processing, 1988.

[98]

李文捷、潘海华等，《基于语料库的中文最长名词短语的自动抽取》；陈力为、袁琦编：《计算语言学进展与应用》，北京：清华大学出版社，1995， pp.119-125。

لي وين جييه، وبان خاي خوا، ((الاستخراج الآلي للتعبيرات الاسمية الطويلة في اللغة الصينية اعتمادًا على ذخيرة لغوية))؛ تشن لي واي، ويوان تشي بيان، ((تطور تطبيقات علم اللغة الحاسوبي))، بكين، دار نشر جامعة تشين خوا، ١٩٩٥، صفحة ١١٩-١٢٥.

[99]

赵军《汉语基本名词短语的识别和结构分析研究》，博士论文，清华大学计算机科学和技术系，1998。

جاو جوين، ((التعبيرات اللغوية الاسمية الأساسية في اللغة الصينية دراسة في التمييز الآلي والتحليل البنائي))، رسالة دكتوراة، قسم علوم الحاسب والتكنولوجيا بجامعة تشين خوا، ١٩٩٨.

[100]

张卫国《三种定语、三类意义及三个槽位》，中国人民大学学报，1996，No.4，pp.97-100。

جانغ وي قوه،"ثلاثة حلول لثلاثة أنوع من المحددات، وثلاثة أنواع من المعنى"، ((المجلة العلمية لجامعة الشعب الصينية))، العدد الرابع، ١٩٩٦، صفحة ٩٧-١٠٠.

[101] Brill, E. Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging, In: Comutational Linguistics, V21. No.4, 1995.

[102] Ramshaw, L., Marcus R. Text chunking using transformation-based leaming, In: Proceedings of the Fourth Wrkshop on Very Large Corpus, 1995, pp.82-94.

[103] Lesk, Michael, Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone. In Proceeding of the 1986 SIGDOC conference, pp. 24-26.

[104] Wilks, Yorick A. and Dan Fass. Preference semantics: A family history. Report MCCS-90-194, Computing Research Laboratory, New MexicoStateUniversity, Las Cruces, NM.

[105] Yarowsky David, Word sense disambiguation using statistical model of Roget's categories trained on large corpora. Proceedings of the 14th International Conference on Computational Linguistics, COLING'92, pp.454-460, Nantes, France, 1992, August.

[106] D.Yarowsky, Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. In: proc 32nd Annual Meeting of Association for Computational Linguistics, 1994, pp.88-95.

[107] Bruce R., A Statistical Method for Word Sense Disambiguation, [Ph.D.Dissertation] USA: New MexicoStateUniversity, 1995, pp.

[108]

梅家驹、竺一鸣、高蕴琦《同义词次林》，上海：上海辞书出版社，1983。

مي جيا جو، وجو يي مينغ، وجاو وين تشي، ((معجم غابة المترادفات))، شنغهاي، دار نشر المعاجم بشنغهاي، ١٩٨٣.

[109] Firth, J.R. 1957.Modes of Meaning.Papers in Linguistics 1934-1951, pp. 190-215, OxfordUniversity Press, Oxford, UK.

المؤلف في سطور[1]

المُؤلِّف الرئيس لهذا الكتاب هو الأستاذ الدكتور خوانغ تشانغ نينغ، أحد أشهر علماء اللغة الحاسوبيين في الصين. وُلِدَ خوانغ تشانغ نينغ عام ١٩٣٧ في مقاطعة جوانغ دونغ، في عام ١٩٦١ تخرج في قسم المحركات الكهربية بجامعة تشينغ خوا ببكين. وفي عام ١٩٨٣ سافر في بعثة إلى جامعة ييل (Yale University) بالولايات المتحدة الأمريكية لمدة عام، وعمل في التدريس لمدة عام بجامعة العلوم والتكنولوجيا بهونج كونج عام ١٩٨٦. وفي عام ١٩٩٩، عمل مديرًا لمركز أبحاث شركة ميكروسوفت في الصين. ويَنْظُر إليه الصينيون باعتباره ممهد الطريق لتوطين علوم المعالجة الآلية للغات الطبيعية في الصين. في سبعينيات القرن العشرين ذاع صيته كأستاذ متميز في جامعة تشينغ خوا عندما أَلَّفَ مُقَرَّر (دوائر الترانزستور)؛ حيث حقق هذا المقرر مبيعات زادت عن مليون نسخة داخل الصين. وبعد ذلك، قَدَّمَ ترجمة لكتابين أحدثا أثرًا كبيرًا في المجتمع العلمي الصيني هما كتاب (مبادئ الذكاء الاصطناعي)، وكتاب (البرمجة باستخدام لغة LISP)، ليملأ بذلك الفراغ العلمي الموجود في تلك الفترة.

وقد وصل خوانغ تشانغ نينغ إلى قمة أخرى من قمم الشهرة العلمية عندما عمل في حقل الدراسات المتعلقة بعلم معالجة اللغات الطبيعية؛ حيث

(١) ملحوظة: الكتاب الأصلي لا يحتوي على تعريف بالمؤلف، ولذلك تم التعريف به في مقدمة المترجم. بالإضافة إلى كتابة هذا التعريف في نبذة مستقلة.

بدأ بتنظيم مجموعات بحثية داخل الصين وخارجها في هذا المجال، ثم تولى رئاسة العديد من المؤتمرات الدولية وهيئات المبرمجين في مجال المعالجة الآلية للغات الطبيعية. ويتولى خوانغ تشانغ نينغ الآن رئاسة تحرير (المجلة العلمية للمعلوماتية الصينية)، بالإضافة إلى كونه مُحكمًا وعضو هيئة تحرير العديد من المجلات العلمية الدولية؛ مثل: (المجلة التخصصية في معالجة المعلومات باللغات الآسيوية ACM) بالولايات المتحدة الأمريكية، و (النشرة الدورية لجمعية معالجة المعلومات باللغة الصينية واللغات الشرقية) بسنغافورة، ومجلة (معالجة اللغات الطبيعية) باليابان، ومجلة (دراسات نقدية حول علم اللغة في الصين) بهونج كونج، ومجلة (علم اللغة الحاسوبي ومعالجة اللغات الصينية) بتايوان[1].

وقد شارك خوانغ تشانغ نينغ في مشروعين بحثيين في إطار الخطة الخمسية السابعة للحكومة الصينية عام ١٩٨٥، هما: مشروع "الفهم الآلي للغات الطبيعية والبرامج البينية بين الإنسان والآلة" وكان ذلك في نطاق مشروعات العصف الذهني، ومشروع "تقنيات الفهم الآلي للنصوص العسكرية" في نطاق المشروعات البحثية المستقبلية للدفاع الوطني، بالإضافة إلى مشروع "نظام للترجمة الآلية بين اللغتين الصينية واليابانية" بالتعاون بين جامعتي تشينغ خوا ونانكين الصينيتين. وقد حصل هذا المشروع على المركز

(١) لمزيد من المعلومات عن المؤلف، انظر الموسوعة الصينية على الرابط التالي:
http://baike.baidu.com/view/1615502.htm

الثاني في مجال تطوير العلوم والتكنولوجيا من لجنة التعليم بالحكومة الصينية.

وقد سبق أن تولى خوانغ تشانغ نينغ رئاسة الفريق البحثي الصيني الذي كُلِّفَ من وزارة الصناعات الإلكترونية الصينية للمشاركة في مشروع الترجمة الآلية متعدد اللغات الذي رصدت له الحكومة اليابانية خمسين مليون دولار أمريكي عام ١٩٩٠ لتطوير نظام للترجمة الآلية بين لغات خمس دول آسيوية هي: اليابان، والصين، وماليزيا، وسنغافورة، وتايلاند.

وفي عام ١٩٩٥ شارك في مشروع بحثي للترجمة الآلية بين اللغتين الإنجليزية والصينية ممثلا عن جامعة تشينغ خوا الصينية بالتعاون مع شركة DEC الأمريكية. كما نجح مع فريقه البحثي في بيع حقوق الملكية الفكرية لعدد من النتائج البحثية لشركتي آي بي إم IBM، وميكروسوفت Microsoft.

لي جوان تزي

تخرجت عام ١٩٨٦ في قسم علوم الحاسب والتكنولوجيا بجامعة شان شي الصينية، وحصلت على الدكتوراه عام ٢٠٠٠ في التخصص ذاته من جامعة تشينغ خوا، وعملت منذ عام ٢٠٠١ بقسم الحاسب بجامعة تشينغ خوا، وتشغل منذ ٢٠٠٤ منصب نائب رئيس مركز بحوث البرمجيات التابع لقسم علوم الحاسب والتكنولوجيا بالجامعة نفسها، كما أنها عضو بجمعية الحاسب الصينية منذ عام ٢٠١٠.

أهم توجهاتها البحثية معالجة المعلومات باللغة الصينية، واكتشاف المعارف وإدارتها في بيئة الإنترنت، شاركت أكثر من مرة كعضو رئيس في مشروعات

الحكومة الصينية في مجال المعلومات؛ حيث عملت في الفترة من ١٩٩٠-١٩٩٢ كمسؤول رئيس عن المشروع البحثي لـــ "التحليل الآلي للجمل في اللغة الصينية"، كما عملت باحثًا رئيسًا في المشروع الرئيس للحكومة الصينية في مجال العلوم الطبيعية عن "النظريات والمنهجيات والأدوات البحثية المستخدمة في علم الذخائر اللغوية"، وذلك في الفترة من ١٩٩٦-١٩٩٩. وحازت عام ١٩٩٤ على جائزة التقدم التكنولوجي من الدرجة الأولى من مقاطعة شان شي.

المترجم في سطور

هشام موسى المالكي

أستاذ الترجمة واللغويات الحاسوبية، بقسم اللغة الصينية، بكلية الألسن – جامعة عين شمس.

حصل على ماجستير الألسن في الترجمة التحريرية عام ١٩٩١.

حصل على دكتوراه الألسن في اللغويات التقابلية بين الصينية والعربية عام ١٩٩٧.

حصل على درجة أستاذ في الترجمة واللغويات الحاسوبية عام ٢٠١١.

شارك في تأسيس قسم اللغة الصينية بكلية الآداب جامعة القاهرة عام ٢٠٠٤.

الإنتاج العلمي البحثي:

قام بتأليف ونشر العديد من الأبحاث في مجال اللغويات التطبيقية، والترجمة، واللغويات الحاسوبية، نذكر منها على سبيل المثال:

١- "إشكاليات تهيئة الذخائر اللغوية وبنائها حاسوبيًا: اللغتان العربية والصينية نموذجًا" (نشر في مجلد المؤتمر الثاني عشر لمعالجة المعلومات بلغات القوميات الأقلية بالصين).

٢- "بناء قاعدة بيانات للتصاحب اللغوي في الصينية والعربية: دراسة إحصائية حاسوبية" (نشر في مجلد أبحاث المؤتمر الدولي التاسع لتعليم اللغة الصينية للأجانب بالصين، وحصل على جائزة البحوث المبتكرة).

٣- "قواعد البيانات الإلكترونية وتطوير صناعة المعاجم المتخصصة ثنائية اللغة".

٤- "إشكاليات ترجمة أسماء الأعلام بين اللغتين الصينية والعربية".

٥- "اتجاه جديد لتعليم الرموز الصينية انطلاقًا من المدخل المنظومي في التعليم والتعلم" (نُشر في مجلد أبحاث المؤتمر الدولي الثامن لتعليم اللغة الصينية للأجانب بالصين).

٦- "الجامعة العربية المفتوحة وتعليم اللغات الأجنبية: اللغة الصينية نموذجًا".

٧- "المدخل المنظومي والوسائط المتعددة في تعليم وتعلم الترجمة الفورية".

قام بتأليف عدد من المقررات الدراسية في اللغويات التطبيقية والترجمة:

١- (محاضرات في الاستماع باللغة الصينية).

٢- (مبادئ الترجمة التحريرية من الصينية إلى العربية).

٣- (الجملة المركبة بين اللغتين الصينية والعربية: دراسة تقابلية).

الإنتاج في مجال الترجمة:

ترجمة كتاب (فن الحرب: النص الأصلي لسون تزي وشروحه المعاصرة)، صدر ضمن أعمال المشروع القومي للترجمة، العدد ٩٦٧، ٢٠٠٥.

وضع الخطة الخمسية للمركز القومي للترجمة على مستوى اللغة الصينية عام ٢٠٠٥.

تنظيم أربع ورش عمل في الترجمة التخصصية لإعداد كوادر من الطلاب بكلية الألسن جامعة عين شمس، نتج عنها أربعة أعداد من مجلة الجسر التي يرأس تحريرها.

تخطيط ورشة للترجمة التخصصية عن اللغة الصينية وتنفيذها، التي انعقدت في المركز القومي للترجمة في العام التدريبي ٢٠١١-٢٠١٢.

تنظيم دورة صيفية مجانية لمدة ثلاثة أشهر لعدد ٣٩ طالبًا وطالبة من قسم اللغة الصينية بكلية الألسن جامعة عين شمس للتدريب على مهارات الترجمة من خلال فريق عمل، ونتج عن الدورة ترجمة كتاب في ٢٣٠ صفحة من اللغة الصينية بعنوان (كيف يربح أبناؤنا في عصر الإنترنت).

تأسيس جماعة رؤية للترجمة من اللغة الصينية من طلاب قسم اللغة الصينية بكلية الآداب جامعة القاهرة عام ٢٠٠٥، وكان أول عمل تمت ترجمته كتاب (قطوف من الحكمة الصينية).

مراجعة كتاب (كيف تجعل أبناءك يحبون الدراسة)، تأليف يانغ شيا، ترجمة مي عاشور، والذي صدر ضمن أعمال المشروع القومي للترجمة، العدد ١٩٢٣، عام ٢٠١١.

الإشراف على إعداد عدد منالمترجمين عن اللغة الصينية من خلال اقتراح سلسلة من الكتب المتخصصة باللغة الصينية، وترشيح المترجمين الجدد وتدريبهم.

ترجمة سلسلة من المقررات الدراسية لتعليم اللغة الصينية للناطقين باللغة العربية، بعنوان (اللغة الصينية في مائة محاضرة).

التصحيح اللغوى : نعيمـــــــــة عاشـــــــــور

الإشـراف الفنـى : حــــسن كامــــل